MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0420
Measure Title: Pain Assessment and Follow-Up
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: NOTE: Specification information in this section is from the 2016 Physician Quality Reporting System Manual. Testing Information is based on the specification in the 2013 (Registry Data) and specification in the 2014 (Claims Data) Physician Quality Reporting System Manual. Specifications from 2013, 2014 and 2016 are included in the attached "NQF Endorsement Measurement Submission Summary Materials"

Note to PFCC Standing Committee: The developer will be provided the opportunity to update their form and clarify the measure specification under consideration during this phase of work. The measure has undergone significant changes since their last endorsement review and a full history is documented. NQF staff have highlighted the sections under consideration.

2014-2016 Specification Description:
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment using a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present

2013 Specification Description (used in Registry Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment through discussion with the patient including the use of a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present

Developer Rationale: This measure addresses a gap in care. There are disparities in care across population groups as outlined in the following statements:

The American Pain Foundation (2009) identified medically underserved populations endure a disproportionate pain burden in all health care settings.

A growing body of research reveals even more extensive gaps in pain assessment and treatment among racial and ethnic populations, with minorities receiving less care for pain than non-Hispanic whites (Green, 2003; Green, 2007; Green et al., 2011; Todd et al., 2004; Todd et al., 2007). Differences in pain care occur across all types of pain (e.g., acute, chronic, cancer-related) and medical settings (e.g., emergency departments and primary care) (Green, 2003; Green, 2007; Todd et al., 2007). Even when income, insurance status and access to health care are accounted for, minorities are still less likely than whites to receive necessary pain treatments (Green, 2003; Green, 2007; Paulson et al., 2007). Black race is associated with neighborhood socio-economic status (SES) and race plays a role in pain outcomes beyond SES (Green, 2012).

Research also shows gender differences in the experience and treatment of pain. Most chronic pain conditions are more prevalent among women; however, women's pain complaints tend to be poorly assessed and undertreated (Green,

2003; Chronic Pain Research Alliance 2011, Weimer 2013). Although women may have higher baseline pain, differences in pain levels may not persist at one month (Peterson, 2012).

"When assessing and treating pain, practitioner sex, race, age, and duration of experience were all significantly associated with pain management decisions. These findings suggest that pain assessment and treatment decisions may be impacted by the health care providers' demographic characteristics, effects which may contribute to pain management disparities."(Bartley et al., 2015).

The aim of this quality measure is to assist eligible providers to identify patients experiencing pain and provide a follow-up plan which addresses the patients' pain in an effort to reduce or eliminate the pain. Ultimately, reducing or eliminating pain will improve a patients' quality of life, minimize the disparities that exist in the assessment and treatment of pain and reduce the cost and utilization of healthcare resources.

Numerator Statement: 2013 Specification Numerator Statement (used in Registry Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment through discussion with the patient including the use of a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present (Testing completed on Registry Data)

2014 and 2016 Numerator Statement (used in Claims Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment using a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present.
Denominator Statement: All visits for patients aged 18 years and older
Denominator Exclusions: : Not Eligible – A patient is not eligible if one or more of the following reason(s) is documented:

Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools

Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

Measure Type: Process
Data Source: Administrative claims, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Jul 31, 2008

Maintenance of Endorsement – Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**

- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary:

- The developer indicated they have updated the evidence since the last endorsement review and stated the following rationale supporting the measure: *Utilization of validated pain assessment tools facilitates the monitoring of the patient's health status and the differentiation of treatment approaches in order to improve the patient's pain level.*
- Three clinical practice guidelines were provided to support the measures: Assessment and Management of Chronic Pain (2013), Adult Acute and Subacute Low Back Pain (2012) and Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability, and Health from the Orthopaedic Section of the American Physical Therapy Association (2012).
- This measure is a process measure that has a more global target population of all adults, two out of the three guidelines cited focus on low back pain. One of the low back pain guidelines is more specific to imaging used in diagnostics versus the pain assessment and follow-up plans

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: See above, the developer submitted a new evidence form. This measure was originally recommended for time-limited endorsement in 2008. The steering Committee that reviewed the measure recommended the changes the developer has made since that time (clarity on standardized assessment, documentation of follow-up plan).

Exception to evidence

Based on the information provided, is there rationale to support this measure with an exception to evidence? As a process measure, the evidence requirement is a systematic assessment and grading of the quantity, quality and consistency of the body of evidence that measured process leads to a desired health outcome. The developers provide clinical guideline recommendations for adult pain and low back pain, and specifically on the importance of assessment. We are specifically looking for evidence that the assessment and documentation of a treatment plan for pain leads to improved outcomes. The lack of systematic assessment of evidence may be an oversight versus the lack of evidence.

Guidance from the Evidence Algorithm

For a process measure, is it based on systematic review and grading of the BODY of empirical evidence where specific focus of the evidence matches what is being measured (box 2): No → is empirical evidence submitted but without systematic review and grading of the evidence (box 7): No → Are there, or could there be , performance measures of a related health outcome or evidence-based intermediate clinical outcome or process (box 10): No → is there evidence of systematic assessment of expert opinion that the benefits outweigh potential harms (box 11): Yes → Does the SC agree that it is okay to hold the providers accountable for performance in the absence of empirical evidence?: if yes – Rate as insufficient evidence with exception; if No – rate as insufficient.

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *Questions specific to the measure information provided on evidence*
  - *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*
  - *Is the evidence directly applicable to the process of care being measured?*
- *For possible exception to the evidence criterion:*
  - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based*

intermediate clinical outcomes, intervention/treatment?

- *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
- *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ High ☐ Moderate ☐ Low ☒ Insufficient

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following summary of performance data from PQRS:
A. Quality Indicator Performance 1/1/2014 through 12/31/2014
1. Total Claims Submitted- 10,555,143
2. Valid Denominator Criteria - 9,515,468/ 90.2% of total
3. Performance Exclusion – 341,159/ 3.5% of valid
4. Measure Performance Rate- 7,627,424 / 9,174,309 83.1%

B. Performance Variation by Eligible Professional 1/1/2014 through 12/31/2014: Describes the variation of measure scores by discrete National Provider Identification (NPI).
• N (# of NPIs) – 59,722
• Mean Measure Score – 81.9%
• Standard Deviation - .35
• Min/Max – 0/100%
• 1st percentile – 0.0%
• 5th percentile – 0.0%
• 10th percentile – 0.0%
• 25th percentile – 90.6%
• 50th percentile – 100.0%

The developer also notes: Reporting for the measure is voluntary and providers who report may not be representative of all eligible professionals. In 2014 only 10.7% of eligible providers reported this measure. Reported performance rates from this group cannot be generalized to the total eligible population

Disparities
Disparities in performance based on race/ethnicity, urban/rural status, gender and age were identified. Analysis of claims from 1/1/2014 through 12/31/2014 reveal statistically significant differences in measure performance between genders and age groups with larger differences observed between urban/rural providers and patient race/ethnic group.

Performance rates by categories:
Rural 87.3%, Urban 81.8% (X2 = 34753.95, N = 9,159,741 p < .0001)
Female 83.7%, Male 82.2 % (X2 = 3424.87, N = 9,174,309 p < .0001)
White 84.2%, Non-white 70.6% (X2 = 85850.38, N = 9,002,090 p < .0001)
Asian 76.2%, Black 68.2%, Hispanic 79.1%, Native 73.6%, White 84.2%, Other 79.6%, Unknown 86.1%(X2 = 95281.16, N = 9,174,309 p < .0001))
Age Under 50 years 80.0%, 50-64 years 80.9%, 65-69 years 85.4%, 70-74 years 84.6%, >=75 81.7% (X2 = 23394.64, N = 9,174,309, p < .0001)

Questions for the Committee:
- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1.a. Evidence to Support Measure Focus

Comments:

**Providing clinical guidelines only supports the premise that assessment and a plan of treatment is important, in that it is necessary but not sufficient to improve pain. The developers did not provide evidence that assessing pain and documenting a plan resumed in improved pain scores, or improved quality of life or function. There is no way of knowing if the plan documented is evidence based or effective. The guidelines sipped are tangentially related, rather than directly related. I am not aware of any studies that either support or refute that better assessment results in improved health outcomes.

**The measure developer sites guidelines that recommend screening for pain and there was a comment as to whether the screening and development of a plan improved patients' outcomes for pain management.
An article published in 2007 questioned the Accuracy of the Pain Numeric Rating Scale as a Screening Test in Primary Care:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2305860/

**The evidence supports the assessment of pain in adults with low back pain. The evidence does suggest that treatment and improvement of pain is a goal worthy of investigation. However, one source notes, "1.Increase the identification of patients who are in the early stages of a serious illness who would benefit from palliative care.
2.Improve the effectiveness and comfort level of primary care clinicians in communicating the necessity and benefits of palliative care with those patients with a serious illness.
3.Improve the assessment of the identified patient's palliative care needs, utilizing the domains of palliative care. 4.Increase the percentage of patients in the early stages of a serious illness who have a care plan identified and/or documented.
5.Improve the ongoing reassessment and adjustment of the patient's plan of care as the condition warrants, utilizing the domains of palliative care.
6.Increase the completion, documentation and ongoing utilization of advance directives for patients with a serious illness."
https://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_palliative_care_guidelines/palliative_care/

A second source concurs that assessment and planning should identify the type and source of chronic pain and the plan should match the finding based on the assessment. They also note the aims as follows:
Aims
1. Improve the function of patients age 18 years and older with chronic pain. (Annotations #2, 14)
2. Improve the assessment and reassessment of patients age 18 years and older with chronic pain diagnosis utilizing the biopsychosocial model. (Annotations #2, 3, 12)
3. Improve the appropriate use of Level I and Level II treatment approaches for patients age 18 years and older with chronic pain. (Annotations #14, 19, 25)
4. Improve the effective use of non-opioid medications in the treatment of patients age 18 years and older with chronic pain. (Annotations #15, 19)
5. Improve the effective use of opioid medications in the treatment of patients age 18 years and older with chronic pain. (Annotations #15, 19)
https://www.icsi.org/_asset/bw798b/ChronicPain.pdf

However, the Faces Pain Scale (FPS) was designed for use in children and does not include instructions on assessing intensity, quality of pain, etc. http://www.iasp-pain.org/Education/Content.aspx?ItemNumber=1519
In addition, the rationale specifically states that the goal is assessment of all types of chronic pain, yet the evidence several discussions limited to the assessment and treatment of chronic low back pain.

Given that the acceptable measures include the faces scale, which is a 1-10 pain scale

1b. Performance Gap

Comments:
**There does appear to be an ongoing performance gap between urban and rural providers, and patient ethnic group. Black patients remain under assessed and treated compared to white patients, with other non-white patients displaying smaller gaps compared to whites.
**The developer supplied data showing variation in results although overall good performance. Since it is a voluntary measure it is possible higher performing groups chose to submit. Only about 10% of eligible providers submitted.
**While the resources, do, support the use of a performance measure related to chronic pain, the measure, as it is proposed, does not assess the outcome of the treatment. From a patient and family centric view of this measure, pain assessment and planning has little value without producing some benefit.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative Claims data

Specifications:

- Satisfactory reporting criteria are met by valid submission of one of six G codes on claims that meet denominator criteria
- The measure is reported via G-codes (numerator and exclusions) and CPT codes (denominator)
- The numerator reporting options are performance met, pain assessment not documented patient not eligible, and pain assessment not documented reason not given (all reported via G-codes)
- This is a process measure and is not risk adjusted

Questions for the Committee :

- *Specific questions on the specifications, codes, definitions, etc.*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Describe any updates to testing
Because of the updates to the specifications over time, and the ability to gather data through PQRS, the developer updated testing to reflect the current measure specifications (use of G-Codes)

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☒ **Data element** ☒ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

Critical Data Element Testing: Quality Insights of Pennsylvania (Quality Insights) oversees the abstraction of 405 randomly generated Medicare Part B claims records for all 74 unique NPIs/eligible professionals who reported one of the G-codes for the measure during the 1/1/2014 – 12/31/2014 time period. Quality Insights requests the medical record documentation from the NPI/eligible professional for the randomly selected encounter date. The documentation is abstracted and a G-code is assigned by two registered nurse (RN) abstractors, one from Quality Insights and one from an independent reviewer contracted with Quality Insights, according to the measure specifications.

Agreement rates between independent reviewers were calculated (inter-rater reliability) as well as the rate of agreement between the numerator code submitted with the claim and an independent reviewer (critical data element validity. See 2b2. Validity testing). Crude agreement, prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals were calculated.

Performance Score: reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed.

Results of reliability testing

Inter-Rater Reliability:
Numerator crude agreement 95.0%
Prevalence adjusted kappa .90 (CI .86 – .94)
Kappa .87 (CI -.81 – .93)

Performance measure score (1/1/2013 – 12/31/2013):

Data source	N	Between-provider variance	Reliability mean	Reliability median	Reliability Std dev	Reliability min/max
Claims	29,398	.105	.994	1.0	.020	.457 - 1.0
Registry	5,639	.214	.996	1.0	.012	.817 – 1.0

Guidance from the Reliability Algorithm

Are specifications precise and complete (box 1): Yes →Was empirical reliability testing conducted (box 2): Yes → Was reliability testing conducted with computed performance measure scores for measured entity (box 4): Yes → Was method described appropriate (box 5): Yes → Based on reliability statistics and scope – what is level of certainty or confidence that the performance measure scores are reliable (box 6): High

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence

- We would like the committee to discuss; while evidence form was submitted and contained clinical recommendations, there may be additional evidence to support this measure that was not submitted. Based on what was on the evidence

form, staff would rate this as "somewhat" met; however, it seems appropriate that a pain assessment would be conducted and follow-up plan documented and this was the recommendation of past committees.

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
Note: the prior measure testing forms were not found thus information is updated in this form.

Describe any updates to validity testing
The developer indicated on their measure checklist that they did not update validity testing, but noted in their testing form that patient level data elements were assessed. This is described below.

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
Quality Insights of Pennsylvania (Quality Insights) oversees the abstraction of 405 randomly generated Medicare Part B claims records for all 74 unique NPIs/eligible professionals who reported one of the G-codes for the measure during the 1/1/2014 – 12/31/2014 time period. Quality Insights requests the medical record documentation from the NPI/eligible professional for the randomly selected encounter date. The documentation is abstracted and a G-code is assigned by two registered nurse (RN) abstractors, one from Quality Insights and one from an independent reviewer contracted with Quality Insights, according to the measure specifications.

Agreement rates between independent reviewers were calculated (inter-rater reliability) as well as the rate of agreement between the numerator code submitted with the claim and an independent reviewer (critical data element validity). Crude agreement, prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals were calculated.

Validity testing results:
Critical data element testing:
Overall Reliability of Claims vs. Independent Review:
Numerator crude agreement 85.9%
Prevalence adjusted kappa .72 (.66 - .79)
Kappa .55 (86% CI .45 - .65)

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Other specific question of the validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- A patient is not eligible if one or more of the following reason(s) is documented:

- Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools
- Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

QIP analyzed 10,555,143 claims submitted for this measure. Of those 9,515,468 (90.2%) met the denominator criteria for patient age and relevant CPT codes as defined in the measure specifications. It was from that pool the sample for reliability testing was drawn. Two independent clinical reviewers abstracted 405 cases from 74 providers to assess validity of exclusion criteria in claims reporting for encounters from 1/1/2014 to 12/31/2014.

3.6 % of the total number of valid claims were reported as exclusions.

Testing of exclusion criteria agreement demonstrated high reliability in measure reporting. Reliability between two independent clinical reviewers was almost perfect with a PAK = .98, (95% CI=.96 - 1.0) and crude agreement= 99.0%; similarly the "gold standard" clinical reviewer vs. claims agreement was almost perfect with a PAK = .98 (99% CI .97 - 1.00), crude agreement=99.2%.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:
Reported provider performance variation (2014):
N – 59,722
Mean – 81.9%
Min – 0.0%, Max – 100.0% Std Deviation .35
50th percentile – 100.0%
25th percentile – 90.6%
10th percentile – 0.0%
1st percentile – 0.0%

- *The overall performance rate reported via claims for the period 1/1/2014 to 12/31/2014 was 83.1%. The average provider performance rate was 81.9%.*
- *Average reported performance rates are above 80% however the need for improvement can be seen for the lowest 10% reporting (10th percentile 0.0%). It should also be noted that the measure is reported voluntarily and those eligible professionals who chose to report may not be representative of the total population of eligible providers.*

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The number of eligible providers reporting the measure is about 10.7% (3.6% in 2010, 4.5% in 2011, 1.8% in 2012, and 7.4% in 2013).

Because reporting is voluntary the reporting population cannot be said to be representative of the total eligible

population. Generalizations to the overall eligible population should not be made.

Guidance from the Reliability Algorithm

Measure specifications consistent with evidence (Box 1): Yes →All relevant potential threats to validity assessed (Box 2): Yes → empirical validity testing using measure as specified (Box 3): Yes → Validity tested at computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes → Based on results and scope of testing and analysis of potential threats, level of certainty/confidence that measure scores are a valid indicator of quality (Box 8): Moderate (some questions about direct evidence support for measure as specified; face validity information not particularly useful, yet exclusion testing and overall validity of measure seemed sound)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 & 2b.1 Specifications

Comments:

**All specifications clear. No risk adjustment since this is a process not an outcome measure. The specifications are consistent with the evidence, in that the guidelines recommend validated tools for assessment. Documentation of plan is not as well defined in the guidelines, with the exception of whether imaging is indicated in radicular pain.

**Specifications are clear and the reliability and validity were assessed.

**Numerous measures are offered as meeting the requirements for a valid and reliable measure. I am somewhat concerned that no evidence is offered about the appropriateness of the measures related to various diagnoses. Many are specific to low back, yet the measure under review does not limit its usability to that population. This raises validity concerns.

2a.2 Reliability Testing

Comments:

**Measure score and data level testing were both performed. Reliability was tested between independent reviewers, and between reviewer and submitter. A sufficient n of encounters where included, from all unique participating providers. Appropriate testing methods were used for a pass/fail measure. Sufficient reliability testing demonstrated.

**Chart reviews were done on a sample of the results submitted and were found to be consistent.

**Unclear given the disparity noted in 2b.1.

2b.2 Validity Testing

Comments:

**Again, assessment and plan for treatment of pain are necessary but not sufficient to improve patient's lives. Agree with past committees that its reasonable to perform these first steps, without which, quality care cannot be provided. Data elements tested against the gold standard only. Adequate scope and entities included for reliability testing, with correct method used.

**The results support whether an assessment and plan were done which are consistent with accepted guidelines. What is less clear is whether this results in a patient centered outcome of either less pain or increased function.

**If I understand correctly, this measure does not evaluate the quality and appropriateness of the tool or the plan. It only assesses whether or not a tool, of any variety, was used, and a plan, also of any variety was created.

2b.3.-2b7. Testing (Related to Potential Threats to Validity)

Comments:

**Exclusion seem reasonable and are sufficiently rare (about 3%.)Exclusions were also reliably identified. Exclusion groups narrow, meaning the vast majority of patients would be included. No patient groups unfairly excluded. The analysis supports that the bottom 10% percentile have lots of room for improvement.

**Exclusions noted along with frequency. Stratification was done to show gaps related to a number of factors.

**In addition, the high floor value and low ceiling suggest that process measure will not be useful in identifying the disparities in chronic pain care it seeks to tease out.

B. Performance Variation by Eligible Professional 1/1/2014 through 12/31/2014: Describes the variation of measure scores by discrete National Provider Identification (NPI).

- N (# of NPIs) – 59,722
- Mean Measure Score – 81.9%
- Standard Deviation - .35
- Min/Max – 0/100%
- 1st percentile – 0.0%
- 5th percentile – 0.0%
- 10th percentile – 0.0%
- 25th percentile – 90.6%
- 50th percentile – 100.0%

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The measure is collected primarily via administrative data (claims), but has an option for medical record abstraction.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**Pain scales of 1-10 almost always generated and captured, but more meaningful scales (VAS, Wong Baker) remain clinically underutilized. A plan of care is rarely documented in any systematic way, and usually involves concerted effort to develop a template that "forces" documentation of a clinically meaningful plan. However, once that process is established, the required data elements can be easily documented in an EHR, and extracted from there with only moderate burden. The upfront investment can be burdensome in other words.
**The groups (10% of eligible) that submitted showed the measure to be feasible. There are concerns that overall feasibility across more clinicians, especially in primary care, could be challenging given the number of things primary care is already expected to do in a given visit. This barrier could be one of the reasons more groups didn't submit this through PQRS.
**No issues noted.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
PQRS

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
The measure is currently in use in the PQRS program; in 2014, there were 573,233 (10.7%) Eligible Professionals who could report NQF# 0420. In 2013, NQF #0420 was the 6th most reported measure within PQRS with 664,929 (7.4%) eligible professionals participating in reporting this measure.

Improvement results
Provider and Patients Statistics for program year 2014 (from "2014 Physician Quality Reporting System Program Monitoring and Evaluation Report"):
Average Performance Rates by Year (PQRS – all reporting methods):
2009 – 97.4%
2010 – 97.3%
2011 – 94.8%
2012 – 86.9%
2013 – 85.7%
2014 - 88.5%

Unexpected findings (positive or negative) during implementation
The developer indicated no unexpected findings

Potential harms
For the overall measure, none noted. For low back pain, it was noted that a standardized, back-specific pain assessment could potentially prevent unwarranted imaging studies.

Feedback :
None; measure was not on the most recent MUC list (2015-6 MAP proceedings)

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4. Usability and Use
Comments:
**Voluntary reporting through PQRS. Continued emphasis on at least assessing and planning to treat and follow progress is undeniably useful, but hope that the process measure eventually becomes an outcome measure, where provider must ensure the plan is evidence based and demonstrating improvement in wellbeing and function.

Documenting these elements might result in more efficient care, if the plan is adequate and results in less resource use, or less cost to the system, with better outcomes.

Criterion 5: Related and Competing Measures

Related or competing measures

0383 : Oncology: Plan of Care for Pain – Medical Oncology and Radiation Oncology (paired with 0384)
0676 : Percent of Residents Who Self-Report Moderate to Severe Pain (Short-Stay)
0677 : Percent of Residents Who Self-Report Moderate to Severe Pain (Long-Stay)
1628 : Patients with Advanced Cancer Screened for Pain at Outpatient Visits
1634 : Hospice and Palliative Care -- Pain Screening
1637 : Hospice and Palliative Care -- Pain Assessment

Harmonization

The developer reports that all measures listed above (and a similar list of measures related, but not endorsed) have not been harmonized, and provided rationale and analysis of differences in measures. Staff review indicates relation to list of measures, and agrees that not competing, mainly due to variations in target population and numerator requirements.

Pre-meeting public and member comments

- We support the pain assessment measure but it is not obvious if any specification for what a "standard" measure of this is—e.g. is a pain scale (what is your pain on a scale from 1-10) sufficient? Also, it is interesting to think about how this gets operationalized in the context of other efforts to try to mitigate overprescribing of opioids. We agree with the need for assessment of pain and a follow-up plan where pain is present, but it is not clear what is acceptable as a follow-up plan—just a prescription and a plan to reevaluate? Referral to pain specialist, PT, etc.?

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0420

Measure Title: Pain Assessment and Follow-Up

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 3/30/2016

All the information in this form is updated from last endorsement of NQF 0420 in September 2011. This NQF evidence form was not in existence in 2010/2011. Evidence continues to support measure focus.

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading

definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

- ☐ Health outcome: Click here to name the health outcome
- ☐ Patient-reported outcome (PRO): Click here to name the PRO

  PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

- ☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

N/A

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

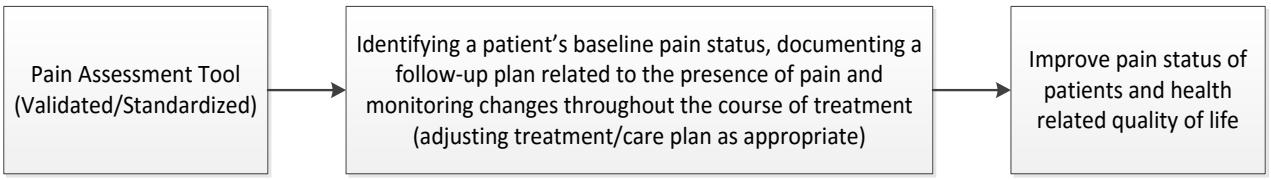
N/A

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Utilization of validated pain assessment tools facilitates the monitoring of the patient's health status and the differentiation of treatment approaches in order to improve the patient's pain level.



1. Assess for the presence of pain using a standardized tool in all patients aged 18 years and older
2. Identification of pain (positive screen) results in the documentation of a follow-up plan related to the presence of pain and the management of it to reduce pain intensity.
3. Follow-up recommendation and intervention strategies for treating pain can lead to decreased level of pain, thus improving the health and well-being of the patient and can help to reduce the use of healthcare resources and/or lost productivity.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1. Hooten, W.M., Timming, R., Belgrade, M., Gaul, .J, Goertz, M., Haake, B., … Walker, N.(2013). Assessment and management of chronic pain. *Institute for Clinical Systems Improvement* (6th ed.). Retrieved from https://www.icsi.org/_asset/bw798b/ChronicPain.pdf

2. Goertz, M., Thorson, D., Bonsell, J., Bonte, B., Campbell,.R, Haake B., …, Timming, R. (2012). Adult Acute and Subacute Low Back Pain. *Institute for Clinical Systems Improvement* (15th ed). Retrieved from https://www.icsi.org/_asset/bjvqrj/LBP.pdf

3. Delitto, A., George, S.Z., Van Dillen, L.R., Whitman, J.M., Sowa, G., Shekelle, P., & Denninger, T.R. (2012). Low back pain. Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability, and Health from the Orthopaedic Section of the American Physical Therapy Association. *Journal of Orthopedic Sports Physical Therapy*, 42(4), A1-A57.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1. ICSI Guideline: Assessment and Management of Chronic Pain (Hooten et al. (2013)) – The assessment and management algorithms are found on pages 1 and 2 of guideline.

A. Assessment Algorithm Annotations (p.12)

Critical First Step: Assessment

Recommendations:

- A clinician should complete an adequate pain assessment on all patients that includes documentation of pain location, intensity, quality, onset/duration/variations/rhythms, manner of expressing pain, pain relief, exacerbation triggers, effects of pain and response to previous treatments.
    - Musculoskeletal assessment - Rasmussen, 2004 [*Low Quality Evidence*]
    - Multidimensional assessment tools – Cleeland, 1994 *[Low Quality Evidence]*, Smith, 1997 [Low Quality Evidence], Galer, 1997 *[Low Quality Evidence],* Savedra, 1989 *[Low Quality Evidence*], VanCleve, 1993 [*Low Quality Evidence),* Penny, 1999 *[Low Quality Evidence]*

General approach to use of pain assessment tools in chronic pain:

- On initial visit, use a multidimensional tool such as the Brief Pain Inventory to obtain a comprehensive picture of the pain experience. The patient should complete this assessment tool before the physician visit.
- With follow-up visits, continue to use a multidimensional pain assessment tool filled out by the patient before seeing the physician.
- Use specific tools such as the Neuropathic Pain Scale (NPS) when appropriate.
- Avoid the use of single-dimensional pain assessment tools in chronic pain except to rate the intensity of specific pain episodes.

(American Pain Society, 2005 [*Low Quality Evidence*]; Herr, 2004 [*Guideline*]; Kaiser Permanente Medical Care Program, 2004 [*Guideline*]; McCaffery, 1999 [Guideline]; Daut, 1983 *[Low Quality Evidence])*

2. ICSI Guideline: Adult Acute and Subacute Low Back Pain (Goertz et al., 2012) **–** Algorithms for Core Treatment of Non-specific Low Back Pain, Red Flags and Radicular Pain are located on pages 1-3 of guideline

A. Recommendations Table for the assessment and treatment of acute and subacute low back pain (p.7)

Topic	Quality of Evidence	Recommendation	Strength of Recommendation	Annotation Number	Relevant References
Education	Moderate	Clinicians should educate patients as an adjunct to other treatment. No standardized form of education is suggested.	Strong	11, 16, 17, 18	*Engers, 2008; Heymans, 2004*

B. Core Treatment of Non-specific Low Back Pain Algorithm Annotations: B. Initial Evaluation and Data Set: Recommendation (p.12)

- Clinicians should not recommend imaging (including computed tomography [CT], magnetic resonance imaging [MRI] and X-ray) for patients with non-specific low back pain *(Strong Recommendation, Moderate Quality Evidence) (Chou 2011; French 2010; Chou 2009b).*

Note: The supportive documentation for this recommendation advises the use of pain assessment tools instead of imaging to influence medical decision-making in the first 6 weeks of onset of non-specific low back pain (p.12).

C. Reevaluation (p. 16)
   - Reevaluation of low back pain should include the following:
     - Pain reassessed with a repeat Visual Analog Scale and Oswestry Disability Questionnaire

3. Low Back Pain: Clinical Practice Guidelines (Delitto et al. (2012))

A. CLINICAL COURSE (p. A2): The clinical course of low back pain can be described as acute, subacute, recurrent, or chronic. Given the high prevalence of recurrent and chronic low back pain and the associated costs, clinicians should place high priority on interventions that prevent (1) recurrences and (2) the transition to chronic low back pain. (Recommendation based on theoretical/foundational evidence.)
B. EXAMINATION – OUTCOME MEASURES (p. A2): Clinicians should use validated self-report questionnaires, such as the Oswestry Disability Index and the Roland-Morris Disability Questionnaire. These tools are useful for identifying a patient's baseline status relative to pain, function, and disability and for monitoring a change in a patient's status throughout the course of treatment. (Recommendation based on strong evidence.)
C. EXAMINATION – ACTIVITY LIMITATION AND PARTICIPATION RESTRICTION MEASURES (p. A2): Clinicians should routinely assess activity limitation and participation restriction through validated performance-based measures. Changes in the patient's level of activity limitation and participation restriction should be monitored with these same measures over the course of treatment. (Recommendation based on expert opinion.)

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

1. ICSI Guideline: Assessment and Management of Chronic Pain (Hooten et al., 2013). See section 1a.4.2 for grade and 1a.4.4 for definition.
2. ICSI Guideline: Adult Acute and Subacute Low Back Pain (Goertz et al., 2012). Strong Recommendation; Moderate Quality Evidence. Definition: see section 1a.4.4
3. Low Back Pain: Clinical Practice Guidelines (Delitto et al. (2012))
   A. CLINICAL COURSE: Recommendation E - (Theoretical/foundational evidence): A preponderance of evidence from animal or cadaver studies, from conceptual models/principles, or from basic science/ bench research supports this conclusion
   B. EXAMINATION – OUTCOME MEASURES**:** Recommendation A - (Strong evidence): A preponderance of level I and/or level II studies support the recommendation
   C. EXAMINATION – ACTIVITY LIMITATION AND PARTICIPATION RESTRICTION MEASURES: Recommendation F - (Expert opinion): Best practice based on the clinical experience of the guideline development team

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1. & 2. ICSI Guidelines Assessment and Management of Chronic Pain Adult Acute and Subacute Low Back Pain use the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system as a method of assessing the quality of evidence and writing recommendations**.** See below for definitions

Grading of Recommendations Assessment, Development and Evaluation (GRADE)

Category	Quality Definitions	Strong Recommendation	Weak Recommendation
High Quality Evidence	Further research is very unlikely to change the work group's confidence in the estimate of effect.	The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients.	The work group recognizes that the evidence, though of high quality, shows a balance between estimates of harms and benefits. The best action will depend on local circumstances, patient values or preferences.
Moderate Quality Evidence	Further research is likely to have an important impact on the work group's confidence in the estimate of effect and may change the estimate.	The work group is confident that the benefits outweigh the risks, but recognizes that the evidence has limitations. Further evidence may impact this recommendation. This is a recommendation that likely applies to most patients.	The work group recognizes that there is a balance between harms and benefit, based on moderate quality evidence, or that there is uncertainty about the estimates of the harms and benefits of the proposed intervention that may be affected by new evidence. Alternative approaches will likely be better for some patients under some circumstances.
Low Quality Evidence	Further research is very likely to have an important impact on the work group's confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.	The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.	The work group recognizes that there is significant uncertainty about the best estimates of benefits and harms.

3. Low Back Pain: Clinical Practice Guidelines (Delitto et al. (2012)) - uses criteria described by the Centre for Evidence-Based Medicine, Oxford for grading the recommendations. See below for definitions.

Oxford Centre for Evidence-Based Medicine

Recommendation Grades

Recommendation A. (Strong evidence): A preponderance of level I and/or level II studies support the recommendation. This must include at least one level I study

Recommendation B. (Moderate evidence): A single high-quality randomized controlled trial or a preponderance of level II studies support the recommendation

Recommendation C. (Weak evidence): A single level II study or a preponderance of level III and IV studies, including statements of consensus by content experts, support the recommendation

Recommendation D. (Conflicting evidence): Higher-quality studies conducted on this topic disagree with respect to their conclusions. The recommendation is based on these conflicting studies

Recommendation E. (Theoretical/foundational evidence): A preponderance of evidence from animal or cadaver studies, from conceptual models/principles, or from basic science/ bench research supports this conclusion

Recommendation F. (Expert Opinion): Best practice based on the clinical experience of the guideline development team

Levels of Evidence:

I. Evidence obtained from high-quality diagnostic studies, prospective studies, or randomized controlled trials
II. Evidence obtained from lesser-quality diagnostic studies, prospective studies, or randomized controlled trials (eg, weaker diagnostic criteria and reference standards, improper randomization, no blinding, <80% follow-up)
III. Case-controlled studies or retrospective studies
IV. Case series
V. Expert Opinion

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

For "Low Back Pain: Clinical Practice Guidelines" (Delitto et al. (2012)) - uses criteria described by the Centre for Evidence-Based Medicine, Oxford for grading the recommendations:

Low Back Pain: Clinical Practice Guidelines Centre for Evidence-Based Medicine, Oxford, United Kingdom URL: (http://www.cebm.net/index.aspx?o=1025)

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

N/A

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

N/A

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

N/A

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1. **ICSI Guideline: Assessment and Management of Chronic Pain** (Hooten et al., 2013) – **"**This guideline discusses the assessment and management of chronic pain. It is intended for primary care clinicians to help with diagnosis and management of primarily four types of biological markers for pain: neuropathic, muscle, inflammatory and mechanical/compressive. Although opioid use is discussed in this guideline, it is not a comprehensive discussion of the usage of opioids in chronic pain."

2. **ICSI Guideline: Adult Acute and Subacute Low Back Pain** (Goertz et al., 2012) – "Adult patients age 18 and over in primary care who have symptoms of low back pain or radiculopathy. The focus is on the acute (pain for up to 7 weeks) and subacute (pain for between 7 and 12 weeks) phases of low back pain. It includes the ongoing management, including indications for spine specialist referral within the first 12 weeks of onset."

3. **Low Back Pain: Clinical Practice Guidelines** (Delitto et al., 2012) "The purpose of these low back pain clinical practice guidelines, in particular, is to describe the peer-reviewed literature and make recommendations related to (1) treatment matched to low back pain subgroup responder categories, (2) treatments that have evidence to prevent recurrence of low back pain, and (3) treatments that have evidence to influence the progression from acute to chronic low back pain and disability."

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1. **ICSI Guideline: Assessment and Management of Chronic Pain** (Hooten et al., 2013). In the guideline, individual study evidence quality was also graded. These evidence grades, when present, are identified in section 1a.4.2

2. **ICSI Guideline: Adult Acute and Subacute Low Back Pain** (Goertz et al., 2012). In the guideline, individual study evidence quality was also graded. These evidence grades, when present, are identified in section 1a.4.2
   Definitions of GRADE: Same as above

3. **Low Back Pain: Clinical Practice Guidelines** (Delitto et al., 2012). Guideline uses criteria by the Centre for Evidence-Based Medicine, Oxford for grading the recommendations. In the guideline, individual study evidence quality was also graded. These evidence grades are identified in section 1a.4.2. Refer to section 1a.4.3 for definitions.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Refer to section 1a.4.2 and 1a.4.3 for grades and definitions.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range:

1. ICSI Guideline: Assessment and Management of Chronic Pain (Hooten et al., 2013) - August 2011-August 2013
2. ICSI Guideline: Adult Acute and Subacute Low Back Pain (Goertz et al., 2012) - May 2011- June 2012.
3. Low Back Pain: Clinical Practice Guidelines (Delitto et al., 2012) - 1966-2010

Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

This information is not provided within the ICSI guideline: Assessment and Management of Chronic Pain, ICSI guideline: Adult Acute and Subacute Low Back Pain or in Low Back Pain: Clinical Practice Guidelines.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

1. This information is not provided within the ICSI Guideline: Assessment and Management of Chronic Pain. The literature search was divided into two stages to identify systematic reviews and randomized controlled trials, meta-analysis and other literature.
2. ICSI Guideline: Adult Acute and Subacute Low Back Pain: The literature search was limited to systematic reviews, meta-analysis and randomized control trials. No further information is provided in guideline.
3. Low Back Pain: Clinical Practice Guidelines: The strength of the body of evidence varies from theoretical/foundational evidence to expert opinion to strong evidence. Definitions for the level of evidence include the following:
   I. Evidence obtained from high-quality diagnostic studies, prospective studies, or randomized controlled trials
   II. Evidence obtained from lesser-quality diagnostic studies, prospective studies, or randomized controlled trials (eg, weaker diagnostic criteria and reference standards, improper randomization, no blinding, <80% follow-up)
   III. Case-controlled studies or retrospective studies
   IV. Case series
   V. Expert Opinion

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1. ICSI Guideline: Assessment and Management of Chronic Pain (Hooten et al. (2013)- Not addressed
2. ICSI Guideline: Adult Acute and Subacute Low Back Pain (Goertz) - Not addressed
3. Low Back Pain: Clinical Practice Guidelines (Delitto et al. (2012)) – Not addressed

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

1. ICSI Guideline: Assessment and Management of Chronic Pain (Hooten et al.,2013)- No harms reported
2. ICSI Guideline: Adult Acute and Subacute Low Back Pain - (Goertz et al., (2012)
   - Harm:
     - No Imaging First Six Weeks with Radicular Pain; Use Core Treatment Plan - Recommendation: Clinicians should not recommend imaging (including CT, MRI or X-ray) for patients in the first six weeks of radicular pain *[Strong Recommendation, Moderate Quality Evidence]*.
       - "Most patients with radiculopathy supported by exam findings consistent with history will recover within several weeks of onset. The majority of disc herniations regress or reabsorb by eight weeks from onset. In the absence of red flags or progressive neurologic deficit there is no evidence that the delaying surgery worsens outcomes. The use of the core treatment plan is recommended. Refer to Annotation #11, Core Treatment Plan. With this in mind, in the face of radiculopathy there is no benefit and there is possible harm in obtaining an MRI prior to six weeks. The exception to this is a progressing neurologic deficit or persistent disabling pain. If the patient has demonstrable leg weakness that is disabling or is worsening, further evaluation with imaging and consultation with a spine specialist would also be indicated" (p.29)
3. Low Back Pain: Clinical Practice Guidelines (Delitto et al., 2012) – No harms reported

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

N/A

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

N/A

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 0420
De.2. **Measure Title:** Pain Assessment and Follow-Up
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services
De.3. **Brief Description of Measure:** NOTE: Specification information in this section is from the 2016 Physician Quality Reporting System Manual. Testing Information is based on the specification in the 2013 (Registry Data) and specification in the 2014 (Claims Data) Physician Quality Reporting System Manual. Specifications from 2013, 2014 and 2016 are included in the attached "NQF Endorsement Measurement Submission Summary Materials"

2014-2016 Specification Description:
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment using a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present

2013 Specification Description (used in Registry Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment through discussion with the patient including the use of a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present
1b.1. **Developer Rationale:** This measure addresses a gap in care. There are disparities in care across population groups as outlined in the following statements:

The American Pain Foundation (2009) identified medically underserved populations endure a disproportionate pain burden in all health care settings.

A growing body of research reveals even more extensive gaps in pain assessment and treatment among racial and ethnic populations, with minorities receiving less care for pain than non-Hispanic whites (Green, 2003; Green, 2007; Green et al., 2011; Todd et al., 2004; Todd et al., 2007). Differences in pain care occur across all types of pain (e.g., acute, chronic, cancer-related) and medical settings (e.g., emergency departments and primary care) (Green, 2003; Green, 2007; Todd et al., 2007). Even when income, insurance status and access to health care are accounted for, minorities are still less likely than whites to receive necessary pain treatments (Green, 2003; Green, 2007; Paulson et al., 2007). Black race is associated with neighborhood socio-economic status (SES) and race plays a role in pain outcomes beyond SES (Green, 2012).

Research also shows gender differences in the experience and treatment of pain. Most chronic pain conditions are more prevalent among women; however, women's pain complaints tend to be poorly assessed and undertreated (Green, 2003; Chronic Pain Research Alliance 2011, Weimer 2013). Although women may have higher baseline pain, differences in pain levels may not persist at one month (Peterson, 2012).

"When assessing and treating pain, practitioner sex, race, age, and duration of experience were all significantly associated with pain management decisions. These findings suggest that pain assessment and treatment decisions may be impacted by the health care providers' demographic characteristics, effects which may contribute to pain management disparities."(Bartley et al., 2015).

The aim of this quality measure is to assist eligible providers to identify patients experiencing pain and provide a follow-up plan which addresses the patients' pain in an effort to reduce or eliminate the pain. Ultimately, reducing or eliminating pain will improve a patients' quality of life, minimize the disparities that exist in the assessment and treatment of pain and reduce the cost and utilization of healthcare resources.

S.4. **Numerator Statement:** 2013 Specification Numerator Statement (used in Registry Data Testing):

Percentage of visits for patients aged 18 years and older with documentation of a pain assessment through discussion with the patient including the use of a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present (Testing completed on Registry Data)

2014 and 2016 Numerator Statement (used in Claims Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment using a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present.
S.7. **Denominator Statement:** All visits for patients aged 18 years and older
S.10. **Denominator Exclusions:** Not Eligible – A patient is not eligible if one or more of the following reason(s) is documented:

Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools

Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

De.1. **Measure Type:** Process
S.23. **Data Source:** Administrative claims, Paper Medical Records
S.26. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Jul 31, 2008 **Most Recent Endorsement Date:** Jul 31, 2008

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** n/a

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_0420_MeasSubm_Evidence_033016.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure addresses a gap in care. There are disparities in care across population groups as outlined in the following statements:

The American Pain Foundation (2009) identified medically underserved populations endure a disproportionate pain burden in all health care settings.

A growing body of research reveals even more extensive gaps in pain assessment and treatment among racial and ethnic populations, with minorities receiving less care for pain than non-Hispanic whites (Green, 2003; Green, 2007; Green et al., 2011; Todd et al., 2004; Todd et al., 2007). Differences in pain care occur across all types of pain (e.g., acute, chronic, cancer-related) and medical settings (e.g., emergency departments and primary care) (Green, 2003; Green, 2007; Todd et al., 2007). Even when income, insurance status and access to health care are accounted for, minorities are still less likely than whites to receive necessary pain treatments (Green, 2003; Green, 2007; Paulson et al., 2007). Black race is associated with neighborhood socio-economic status (SES) and race plays a role in pain outcomes beyond SES (Green, 2012).

Research also shows gender differences in the experience and treatment of pain. Most chronic pain conditions are more prevalent among women; however, women's pain complaints tend to be poorly assessed and undertreated (Green, 2003; Chronic Pain Research Alliance 2011, Weimer 2013). Although women may have higher baseline pain, differences in pain levels may not persist at one month (Peterson, 2012).

"When assessing and treating pain, practitioner sex, race, age, and duration of experience were all significantly associated with pain management decisions. These findings suggest that pain assessment and treatment decisions may be impacted by the health care providers' demographic characteristics, effects which may contribute to pain management disparities."(Bartley et al., 2015).

The aim of this quality measure is to assist eligible providers to identify patients experiencing pain and provide a follow-up plan which addresses the patients' pain in an effort to reduce or eliminate the pain. Ultimately, reducing or eliminating pain will improve a patients' quality of life, minimize the disparities that exist in the assessment and treatment of pain and reduce the cost and utilization of healthcare resources.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This PQRS measure is designed to encourage and improve the documentation and reporting of standardized pain assessments. It is scored as a simple count of valid submissions on payment claims in the time frame where Part B Medicare claims were available for analysis.

The measure is constructed so that a performance score can be easily derived by dividing the number of claims with codes indicating that the recommended processes were followed (or that the patient was ineligible) by the total number of numerator G codes submitted.

2014 Performance Scores: Claims data consists of all Medicare Part B claims submitted from 1/1/2014 to 12/31/2014 with one of the numerator G codes for this measure. The numerator G code submissions are voluntary and providers who report may not be representative of all eligible professionals. Performance rates cannot be generalized to the population.

A. Quality Indicator Performance 1/1/2014 through 12/31/2014
1. Total Claims Submitted- 10,555,143
2. Valid Denominator Criteria - 9,515,468/ 90.2% of total
3. Performance Exclusion – 341,159/ 3.5% of valid
4. Measure Performance Rate- 7,627,424 / 9,174,309 83.1%

B. Performance Variation by Eligible Professional 1/1/2014 through 12/31/2014: Describes the variation of measure scores by discrete National Provider Identification (NPI).
- N (# of NPIs) – 59,722
- Mean Measure Score – 81.9%
- Standard Deviation - .35
- Min/Max – 0/100%
- 1st percentile – 0.0%
- 5th percentile – 0.0%
- 10th percentile – 0.0%
- 25th percentile – 90.6%
- 50th percentile – 100.0%

Performance scores for the majority of reporting providers skew high (90.6% at the 25th percentile) but drop off sharply for the below the 25th percentile (0% at the 10th percentile). As the eligible provider pool has expanded average performance rates decreased (97.4% in 2009, 88.5% in 2014).

Reporting for the measure is voluntary and providers who report may not be representative of all eligible professionals. In 2014 only 10.7% of eligible providers reported this measure. Reported performance rates from this group cannot be generalized to the total eligible population

Provider and Patients Statistics for program year 2014 (from "2014 Physician Quality Reporting System Program Monitoring and Evaluation Report"):
Average Performance Rates by Year (PQRS – all reporting methods):
2009 – 97.4%
2010 – 97.3%
2011 – 94.8%
2012 – 86.9%
2013 – 85.7%
2014 - 88.5%

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
n/a

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Disparities in performance based on race/ethnicity, urban/rural status, gender and age were identified. Analysis of claims from 1/1/2014 through 12/31/2014 reveal statistically significant differences in measure performance between genders and age groups with larger differences observed between urban/rural providers and patient race/ethnic group.

Performance rates by categories:
Rural 87.3%, Urban 81.8% (X2 = 34753.95, N = 9,159,741 p < .0001)
Female 83.7%, Male 82.2 % (X2 = 3424.87, N = 9,174,309 p < .0001)
White 84.2%, Non-white 70.6% (X2 = 85850.38, N = 9,002,090 p < .0001)
Asian 76.2%, Black 68.2%, Hispanic 79.1%, Native 73.6%, White 84.2%, Other 79.6%, Unknown 86.1%(X2 = 95281.16, N = 9,174,309 p < .0001))
Age Under 50 years 80.0%, 50-64 years 80.9%, 65-69 years 85.4%, 70-74 years 84.6%, >=75 81.7% (X2 = 23394.64, N = 9,174,309, p < .0001)

Refer to section IV. Analysis of Claims Data in attached "NQF Endorsement Measurement Submission Summary Materials" document

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
n/a

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
The American Pain Foundation (2009) identified pertinent facts related to the impact of pain as follows:
• Uncontrolled pain is a leading cause of disability and diminishes quality of life for patients, survivors, and their loved ones. It interferes with all aspects of daily activity, including sleep, work, social and sexual relations.
• Under-treated pain drives up costs – estimated at $100 billion annually in healthcare expenses, lost income, and lost productivity–

extending length of hospital stays, as well as increasing emergency room trips and unplanned clinic visits.
• Medically underserved populations endure a disproportionate pain burden in all health care settings
• Disparities exist among racial and ethnic minorities in pain perception, assessment, and treatment for all types of pain, whether chronic or acute.

The Institute Of Medicine's (IOM) Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education and Research (2011) report suggests that chronic pain rates will continue to increase as a result of:
• More Americans will experience a disease in which chronic pain is associated (diabetes, cardiovascular disease, etc.)
• Increase in obesity which is associated with chronic conditions that have painful symptoms
• Progress in lifesaving techniques for catastrophic injuries for people who would have previously died leads to a group of young people at risk for lifelong chronic pain
• Surgical patients are at risk for acute and chronic pain
• The public has a better understanding of chronic pain syndromes and new treatments and therefore may seek help when they may not have sought help in the past

Gaskin and Richard (2012) studied the economic costs of pain in the United States estimates and reported the national cost of pain ranged from $560 to $635 billion, exceeding the annual costs of heart disease, cancer and diabetes. This study also reported chronic pain affects approximately 100 million adults in the USA. Chronic pain impacts the working lives of those affected as well as Independent Activities of Daily Living (IADLs), sleep and the family as noted by Prefontaine and Rochette (2013). Low back pain and neck pain are two of the diseases with the largest number of years lived with a disability (YLDs) in 2010 as reported by The State of US Health, 1990-2010, Burden of Diseases, Injuries, and Risk Factors (Murray et al., 2013). Inflation adjusted ($2010) biennial expenditures on ambulatory services for chronic back pain increased by 129% from $15.6 billion in 2000-2001 to $35.7 billion in 2006-2007 (Smith, 2013). It is clear the enormous pain-related costs, in both dollars and quality of life, represent a great challenge and an opportunity in terms of improving the quality and cost-effectiveness of care.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

American Pain Foundation (2009). Pain resource guide: Getting the help you need. Retrieved from http://www.peacehealthlabs.org/GeneralPurposeDocuments/Pain%20Resource%20Guide.pdf

Gaskin, D. and Richard, P. (2012). The Economic Costs of Pain in the United States. The Journal of Pain, 13(8), 715-724.

Institute of Medicine (2011). A blueprint for transforming prevention, care, education, and research. Relieving pain in america (269-276). Washington, DC: The National Academies Press. Retrieved from: http://www.nap.edu/download.php?record_id=13172#

Murray, C.J., Abraham, J., Ali, M.K., Alvarado, M., Atkinson, C., Baddour, L.M….Lopez, A.D. (2013). The State of US Health, 1990-2010, Burden of Diseases, Injuries, and Risk Factors. JAMA; 310(6), 591-608. doi:10.1001/jama.2013.13805

Prefontaine, K. & Rochette, A. (2013). A literature review on chronic pain: the daily overcoming of a complex problem. British Journal of Occupational Therapy, 76(6), 280-286. DOI: 10.4276/030802213X13706169932905

Smith, M., Davis, M.A., Stano, M., &, Whedon, J. M. (2013). Aging baby boomers and the rising cost of chronic back pain: secular trend analysis of longitudinal medical expenditures panel survey data for years 2000 to 2007. Journal of Manipulative and Physiological Therapeutics, 36(1), 1-9.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
n/a

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across

organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*
Musculoskeletal, Musculoskeletal : Low Back Pain, Prevention : Screening

De.6. Cross Cutting Areas *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Functional Status, Prevention, Prevention : Screening

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.cms.gov/apps/ama/license.asp?file=/PQRS/Downloads/2016_PQRS_IndMeasuresSpecs_ClaimsRegistry_010716.zip
https://www.cms.gov/apps/ama/license.asp?file=/PQRS/Downloads/2016_PQRS_IndivMeasures_SingleSource_12182015.xlsx

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Data_Dictionary_033016.xlsx

S.3. For endorsement maintenance, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
2011 Measure Specification: The Instructions were updated to reflect the term “eligible professional” in place of ‘non-MD/DO clinicians’; the numerator statement now includes the word “therapy” as stated in the title, for consistency; definition of “Standardized Tool” was updated to include examples of standardized tools; updated description of G-Codes by substituting the word “therapy” for the word ‘treatment.”

2012 Measure Specification: the title was updated from “Pain Assessment Prior to Initiation of Patient Therapy and Follow-up” to “Pain Assessment and Follow-Up” to avoid confusion regarding the term “prior to the initiation of therapy;” minor language changes to the Description, Numerator and Instructions; added definition of Pain Assessment; updated Definition of Not Eligible, Standardized Tool, Follow-Up Plan and Not Eligible; deleted definition of Qualifying Visit; added Wellness codes G0402, G0438 and G0439 HCPCS code G0101 (cervical or vaginal cancer screening; pelvic and clinical breast examination) and ‘office or outpatient visit for the evaluation of new or established patient’ codes to Denominator Coding to allow a broader base of providers to report; deleted Denominator CPT Code 99211, this is a five minute office or outpatient visit; replaced Numerator Option codes G-Code G8440, G8508, and G8441 with G8730, G8731, and G8732 which contained more specific descriptions of the quality action performed.

2013 Measure Specification: Minor language changes to Description, added clarifying language to the Instructions linking the follow-up plan to the presence of pain; minor language change to Denominator Statement; added denominator codes for treatment of speech, language, voice, communication, and/or auditory processing disorder, treatment of swallowing dysfunction and/or oral function for feeding and a code for development of cognitive skills to improve attention, memory, and problem solving to allow eligible provider reporting; added quality action numerator code G8939 - Pain assessment documented, follow-up plan not documented, patient not eligible/appropriate for improved reporting; updated psychiatric diagnostic evaluation codes; minor language changes to Numerator Definitions including the removal of “patient refuses to participate” and ‘diagnosis/condition/illness is not situationally related to pain” from the definition of Not Eligible; G-code description language added for ease of reporting and minor language changes to the G-code definitions which do not change the intent of the quality action code.

2014 Measure Specification: Updated description by removing the phrase ‘through discussion with the patient’; provided additional example of a follow-up in the Instructions; added ophthalmological, physical therapy, occupational therapy, dental and neuropsychological testing CPT codes to the denominator coding to broaden eligible provider reporting; added Numerator Note to assist providers with the documentation of the use of a standardized pain assessment tool and included an exception to this documentation; updated Numerator Definitions of Pain Assessment and Follow-Up; all G-code definitions updated by providing more detail.

2015 Measure Specification: Addition of health and behavior assessment denominator CPT code, 96151.

2016 Measure Specification: Updated National Quality Strategy Domain to "Communication and Care Coordination".

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
2013 Specification Numerator Statement (used in Registry Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment through discussion with the patient including the use of a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present (Testing completed on Registry Data)

2014 and 2016 Numerator Statement (used in Claims Data Testing):
Percentage of visits for patients aged 18 years and older with documentation of a pain assessment using a standardized tool(s) on each visit AND documentation of a follow-up plan when pain is present.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
This measure is to be reported for each visit occurring during the reporting period for patients seen during the reporting period. The reporting period is 12 months from January 1st to December 31st.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
2016 Numerator Details (Note: 2013 and 2014 Numerator Details are similar with minor language edits):
Definitions:
Pain Assessment – Documentation of a clinical assessment for the presence or absence of pain using a standardized tool is required. A multi-dimensional clinical assessment of pain using a standardized tool may include characteristics of pain; such as: location, intensity, description, and onset/duration.

Standardized Tool – An assessment tool that has been appropriately normed and validated for the population in which it is used. Examples of tools for pain assessment, include, but are not limited to: Brief Pain Inventory (BPI), Faces Pain Scale (FPS), McGill Pain Questionnaire (MPQ), Multidimensional Pain Inventory (MPI), Neuropathic Pain Scale (NPS), Numeric Rating Scale (NRS), Oswestry Disability Index (ODI), Roland Morris Disability Questionnaire (RMDQ), Verbal Descriptor Scale (VDS), Verbal Numeric Rating Scale (VNRS) and Visual Analog Scale (VAS).

Follow-Up Plan – A documented outline of care for a positive pain assessment is required. This must include a planned follow-up appointment or a referral, a notification to other care providers as applicable OR indicate the initial treatment plan is still in effect. These plans may include pharmacologic and/or educational interventions.

Not Eligible – A patient is not eligible if one or more of the following reason(s) is documented:
• Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools
• Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

NUMERATOR NOTE: The standardized tool used to assess the patient's pain must be documented in the medical record (exception: A provider may use a fraction such as 5/10 for Numeric Rating Scale without documenting this actual tool name when assessing pain for intensity).

G-codes are defined as Quality Data Codes (QDCs), which are subset of HCPCs II codes. QDCs are non-billable codes that providers will use to delineate their clinical quality actions, which are submitted with Medicare Part B Claims. There are 6 G-code options for this measure.

Numerator Quality-Data Coding Options for Reporting Satisfactorily:
Pain Assessment Documented as Positive AND Follow-Up Plan Documented
(One quality-data code [G8730 or G8731] is required on the claim form to submit this numerator option)
Performance Met: G8730: Pain assessment documented as positive using a standardized tool AND a follow-up plan is documented
OR
Pain Assessment Documented as Negative, No Follow-Up Plan Required
Performance Met: G8731: Pain assessment using a standardized tool is documented as negative, no follow-up plan required

OR

Pain Assessment not Documented Patient not Eligible
(One quality-data code [G8442 or G8939] is required on the claim form to submit this numerator option)
Other Performance Exclusion: G8442: Pain assessment NOT documented as being performed, documentation the patient is not eligible for a pain assessment using a standardized tool
OR
Pain Assessment Documented as Positive, Follow-Up Plan not Documented, Patient not Eligible
Other Performance Exclusion: G8939: Pain assessment documented as positive, follow-up plan not documented, documentation the patient is not eligible

OR

Pain Assessment not Documented, Reason not Given
(One quality-data code [G8732 or G8509] is required on the claim form to submit this numerator option)
Performance Not Met: G8732: No documentation of pain assessment, reason not given
OR
Pain Assessment Documented as Positive, Follow-Up Plan not Documented, Reason not Given
Performance Not Met: G8509: Pain assessment documented as positive using a standardized tool, follow-up plan not documented, reason not.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All visits for patients aged 18 years and older

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Lists of individual codes with descriptors for the 2013, 2014, and 2016 measure specifications are provided in an Excel file at S.2b

2013 Specification (used in Registry Data Testing):
Denominator Criteria (Eligible Cases): Patient encounter during the reporting period (CPT or HCPCS): 90791, 90792, 92507, 92508, 92526, 96116, 96150, 97001, 97003, 97532, 98940, 98941, 98942, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, G0101, G0402, G0438, G0439

2014 Specification (used in Claims Data Testing):
Denominator Criteria (Eligible Cases): Patient encounter during the reporting period (CPT or HCPCS): 90791, 90792, 92002, 92004, 92012, 92014, 92507, 92508, 92526, 96116, 96118, 96150, 97001, 97002, 97003, 97004, 97532, 98940, 98941, 98942, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, D7140, D7210, G0101, G0402, G0438, G0439
(Denominator codes for ophthalmological, physical therapy, occupational therapy, dental and neuropsychological testing were added: CPT codes 92002, 92004, 92012, 92014, D7140, D7210, 97002, 97004 and 96118)

2016 Specification:
Denominator Criteria (Eligible Cases): Patient encounter during the reporting period (CPT or HCPCS): 90791, 90792, 92002, 92004, 92012, 92014, 92507, 92508, 92526, 96116, 96118, 96150, 96151, 97001, 97002, 97003, 97004, 97532, 98940, 98941, 98942,

99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, D7140, D7210, G0101, G0402, G0438, G0439

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Not Eligible – A patient is not eligible if one or more of the following reason(s) is documented:

Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools

Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Pain Assessment not Documented Patient not Eligible
(One quality-data code [G8442 or G8939] is required on the claim form to submit this numerator option)
Other Performance Exclusion: G8442: Pain assessment NOT documented as being performed, documentation the patient is not eligible for a pain assessment using a standardized tool
OR
Pain Assessment Documented as Positive, Follow-Up Plan not Documented, Patient not Eligible
Other Performance Exclusion: G8939: Pain assessment documented as positive, follow-up plan not documented, documentation the patient is not eligible

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
All eligible patients are subject to the same numerator criteria

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

Satisfactory reporting criteria are met by valid submission of one of six G codes on claims that meet denominator criteria.
A rate of quality performance is calculated by dividing the number of records with G codes indicating that the quality actions were performed or that the patient was not eligible by total number of valid G code submissions.

THIS SECTION PROVIDES DEFINITIONS & FORMULAS FOR THE NUMERATOR (A), TOTAL DENOMINATOR POPULATION (TDP), DENOMINATOR EXCLUSIONS (B) CALCUATION & PERFORMANCE DENOMINATOR (PD) CALCULATION.

NUMERATOR (A): HCPCS Clinical Quality Codes G8730, G8731

TOTAL DENOMINATOR POPULATION (TDP): Patient aged 18 years and older on the date of the encounter of the 12-month reporting period, with denominator defined encounter codes & Medicare Part B Claims reported HCPCS Clinical Quality Codes G8730, G8731, G8442, G8939, G8732, G8509

DENONINATOR EXCLUSION (B): HCPCS Clinical Quality Code G8442, G8939

DENOMINATOR EXCLUSION CALCULATION: Denominator Exclusion (B): # of patients with valid exclusions # G8442+G8939 / # TDP

PERFORMANCE DENOMINATOR CALCULATION: Performance Denominator (B): Patients meeting criteria for performance denominator calculation # A / (# TDP - # B)

(Refer to section V. Measure Logic Flow Diagram for Performance Rate Calculation in attached "NQF Endorsement Measurement Submission Summary Materials" Document)

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
n/a

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
n/a

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
n/a

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data source is the patient medical record. Medicare Part B claims data and registry data is provided for test purposes.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Ambulatory Care : Outpatient Rehabilitation, Behavioral Health/Psychiatric : Outpatient
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
n/a

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_0420_Testing_Attachment_033016.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0420
Measure Title: Pain Assessment and Follow-Up
Date of Submission: 3/30/2016
All the information in this form is updated from last endorsement of NQF 0420 in September 2011. This NQF testing form was not in existence in 2010/2011. Testing continues to support measure specification.

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

2014 Part B Medicare claims data for HCPCS codes G8730, G8731, G8442, G8939, G8732, G8509.

2013 PQRS Administrative Data for claims and registry

1.3. What are the dates of the data used in testing? Registry/Claims: 1/1/2013 – 12/31/2013, Claims: 1/1/2014 – 12/31/2014

Part B Medicare claims data for encounters from 1/1/2014 to 12/31/2014 were analyzed for performance gaps and variation.

Performance data aggregated at the provider level from PQRS Administrative Data for claims and registry for encounters from 1/1/2013 to 12/31/2013 were analyzed for signal to noise reliability.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice

☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Data element reliability/validity sample (1/1/2014 – 12/31/2014):
A total of 59,722 unique NPIs reported the measure on 10,555,143 claims.

NPIs that had fewer than ten claims were removed from the dataset. A simple random sample of 160 NPIs was drawn from 46,001 remaining NPIs in the claims database. The records were then stratified by the business location address listed in the NPI registry so that the maximum number of records from each business location was limited to 10 records. This limitation was set so that the providers would not see this task as too burdensome and would be more likely to send in their records. The resulting sample was comprised of 761 claims.

Providers were mailed a letter requesting that they provide the documentation to support the assignment of the numerator code that they had submitted on the claim.

Documentation for 405 claims from 74 providers was received and reviewed.
Records Requested/Returned/Reviewed 761/416/405
Providers Requested/Returned/Reviewed 160/75/74
Provider response rate 46.9%

Performance score reliability data (1/1/2013 – 12/31/2013):
29,398 providers reporting via claims with an average of 167 cases per provider.

5,639 providers reporting via registry with an average of 197 cases per provider.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Data element reliability/validity sample (1/1/2014-12/31/2014):
Description of the population reporting the measure via claims:
Claims with Valid Denominator Criteria: 9,515,468/10,555,143 (90.2%)
3.6% were reported as performance exclusions with a total reported performance rate of 83.1%.

76.5% Urban
23.6% Rural

61.2% Female
38.8% Male

92.2% Non-underserved
7.8% Underserved (racial/ethnic minority)

0.8% Asian
5.6% Black
0.9% Hispanic
0.3% Native
90.5% White
0.9% Other
0.9% Unknown

4.8% Under 50
10.6% Aged 50-64
26.2% Aged 65 – 69
22.3% Aged 70 – 74
36.2% Aged 75

Performance score reliability data (1/1/2013-12/31/2013):
Total # of cases:
Claims: 5,004,383
Registry: 1,125,002

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Date element validity/reliability assessed with Part B Medicare claims with patient level detail from 1/1/2014 – 12/31/2014.

Performance score reliability was assessed using provider level performance data reported for PQRS for 2013.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
n/a

2a2. RELIABILITY TESTING
Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Critical data element testing:
Quality Insights of Pennsylvania (Quality Insights) oversees the abstraction of 405 randomly generated Medicare Part B claims records for all 74 unique NPIs/eligible professionals who reported one of the G-codes

for the measure during the 1/1/2014 – 12/31/2014 time period. Quality Insights requests the medical record documentation from the NPI/eligible professional for the randomly selected encounter date. The documentation is abstracted and a G-code is assigned by two registered nurse (RN) abstractors, one from Quality Insights and one from an independent reviewer contracted with Quality Insights, according to the measure specifications.

Agreement rates between independent reviewers were calculated (inter-rater reliability) as well as the rate of agreement between the numerator code submitted with the claim and an independent reviewer (critical data element validity. See 2b2. Validity testing). Crude agreement, prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals were calculated.

Cohen´s kappa represents chance-corrected proportional agreement. High prevalence of responses in a small number of cells is known to produce unexpected results known as the "kappa paradox." When the prevalence of a rating in the population is very high or low the value of kappa may indicate poor reliability even with a high observed proportion of agreement. In some cases, PAK is shown to provide an additional interpretation of agreement when the prevalence of responses is concentrated in a small number of cells.
See also 2b2. Validity testing

Performance measure score:
Reliability was calculated according to the methods outlined in a technical report prepared by J.L. Adams titled "The Reliability of Provider Profiling: A Tutorial" (RAND Corporation, TR-653-NCQA, 2009). In this context, reliability represents the ability of a measure to confidently distinguish the performance of one physician from another. As discussed in the report: "Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error."

According to this approach, reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Critical data element testing:
Inter-Rater Reliability:
Numerator crude agreement 95.0%
Prevalence adjusted kappa .90 (CI .86 – .94)
Kappa .87 (CI -.81 – .93)

See also 2b2. Validity testing.

Performance measure score (1/1/2013 – 12/31/2013):

Data source	N	Between-provider variance	Reliability mean	Reliability median	Reliability Std dev	Reliability min/max
Claims	29,398	.105	.994	1.0	.020	.457 - 1.0
Registry	5,639	.214	.996	1.0	.012	.817 – 1.0

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Critical data element testing:
Inter-rater reliability testing indicates high agreement.

Landis and Koch (1977) have proposed the following as standards for strength of agreement for the kappa coefficient: [less than or equal to] O=poor, .01 -.20=slight, .21 -.40=fair, .41.- 60=moderate, .61-.80=substantial and .81-1 =almost perfect (high). These categories are informal.
See also 2b2. Validity testing.

Performance measure score:
Provider-specific reliability demonstrates a sufficient level of reliability to detect real difference in performance scores.

In general, reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers) whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities.

There is not a clear cut-off for minimum reliability level. Values above 0.7, however, are considered sufficient to see differences between some physicians (or clinics) and the mean, and values above 0.9 are considered sufficient to see differences between pairs of physicians (see RAND tutorial, 2009).

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☐ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Critical data element testing:
Quality Insights of Pennsylvania (Quality Insights) oversees the abstraction of 405 randomly generated Medicare Part B claims records for all 74 unique NPIs/eligible professionals who reported one of the G-codes for the measure during the 1/1/2014 – 12/31/2014 time period. Quality Insights requests the medical record documentation from the NPI/eligible professional for the randomly selected encounter date. The documentation is abstracted and a G-code is assigned by two registered nurse (RN) abstractors, one from Quality Insights and one from an independent reviewer contracted with Quality Insights, according to the measure specifications.

Agreement rates between independent reviewers were calculated (inter-rater reliability) as well as the rate of agreement between the numerator code submitted with the claim and an independent reviewer (critical data element validity). Crude agreement, prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals were calculated.

Face validity:
Quality Insights of Pennsylvania conducts an Environmental Scan to evaluate the most current research and evidence-based guidelines. The TEP, composed of subject matter specialists and experts with technical measure expertise evaluates the results of the review and provides recommendations based on the scientific merits of the evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE). The TEP also reviews and establishes the measure´s ability to capture what it is designed to capture using a consensus process.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Critical data element testing:
Overall Reliability of Claims vs. Independent Review:
Numerator crude agreement 85.9%
Prevalence adjusted kappa .72 (.66 - .79)
Kappa .55 (86% CI .45 - .65)

Face validity:
N/A

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Critical data element testing:
There is substantial agreement between claims reporting and independent reviewer.

Landis and Koch (1977) have proposed the following as standards for strength of agreement for the kappa coefficient: [less than or equal to] O=poor, .01 -.20=slight, .21 -.40=fair, .41- .60=moderate, .61-.80=substantial and .81-1 =almost perfect (high). These categories are informal.

Face Validity:
Based on the process of multiple stakeholder input, expert panel discussion and public comment, face and content validity of CMS/Quality Insights measures can be assumed to be established.

_____
2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

QIP analyzed 10,555,143 claims submitted for this measure. Of those 9,515,468 (90.2%) met the denominator criteria for patient age and relevant CPT codes as defined in the measure specifications. It was from that pool the sample for reliability testing was drawn. Two independent clinical reviewers abstracted 405 cases from 74 providers to assess validity of exclusion criteria in claims reporting for encounters from 1/1/2014 to 12/31/2014.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

3.6 % of the total number of valid claims were reported as exclusions.

Testing of exclusion criteria agreement demonstrated high reliability in measure reporting. Reliability between two independent clinical reviewers was almost perfect with a PAK = .98, (95% CI=.96 - 1.0) and crude agreement= 99.0%; similarly the "gold standard" clinical reviewer vs. claims agreement was almost perfect with a PAK = .98 (99% CI .97 -1.00), crude agreement=99.2%.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Instances of reported exclusions were relatively small (3.6%) of the entire reported population and include:

- Severe mental and/or physical incapacity where the person is unable to express himself/herself in a manner understood by others. For example, cases where pain cannot be accurately assessed through use of nationally recognized standardized pain assessment tools
- Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status

Gold standard agreement with claims as well as agreement between two independent reviewers indicates almost perfect agreement.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

n/a

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

n/a

2b4.4a. What were the statistical results of the analyses used to select risk factors?

n/a

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

n/a

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
n/a

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:** n/a

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): n/a

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves: n/a

2b4.9. Results of Risk Stratification Analysis: n/a

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
n/a

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
n/a

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

This PQRS measure is designed to encourage and improve the documentation and reporting of a pain assessment using a standardized tool and a follow-up plan if pain present. Performance rates are derived by dividing the number of claims with codes indicating that the recommended processes were followed (or that the patient was ineligible) by the total number of numerator reporting codes submitted.

Variation in performance rates were described by measures of central tendency, variation and percentile rankings. Chi-square was used to test for significant differences between expected and observed performance scores for various populations based on demographic traits.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Reported provider performance variation (2014):
N – 59,722
Mean – 81.9%

Min – 0.0%, Max – 100.0% Std Deviation .35
50th percentile – 100.0%
25th percentile – 90.6%
10th percentile – 0.0%
1st percentile – 0.0%

The overall performance rate reported via claims for the period 1/1/2014 to 12/31/2014 was 83.1%. The average provider performance rate was 81.9%.

Performance results by population groups:
Rural: 87.3% (n=2,156,781) Urban: 81.8% (n=7,002,960) (X^2 = 34753.94, p < .0001)
Female: 83.7% (n=5,613,407) Male: 82.2 % (n=3,560,902) (X^2 = 3424.87, p < .0001)
White: 84.2% (n=8,302,925) Non-white: 70.6% (n=699,165) (X^2 = 85850.38, p < .0001)
Asian: 76.2% (n=73,065) Black: 68.2% (n=513,909) Hispanic: 79.1% (n=82,542) Native: 73.6% (n=29,649)
White: 84.2% (n=8,302,925) Other: 79.6% (n=86,090) Unknown: 86.1% (n=86,129) (X^2 = 95002.59, p < .0001))
Age Under 50 years: 80.0% (n=436,357) 50-64 years: 80.9% (n=971,945) 65-69 years: 85.4% (n=2,404,142)
70-74 years: 84.6% (n=2,043,705) >=75: 81.7% (n=3,318,160) (X^2 = 23394.64, p < .0001)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Disparities in performance based on age, race/ethnicity, gender, urban/rural status, etc. can be identified if present.

Analysis of 2014 claims reveals a statistically significant difference in measure performance in relation to the provider´s rural/urban designation as well as patient gender, race and age group.

Average reported performance rates are above 80% however the need for improvement can be seen for the lowest 10% reporting (10^{th} percentile 0.0%). It should also be noted that the measure is reported voluntarily and those eligible professionals who chose to report may not be representative of the total population of eligible providers.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
n/a

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
n/a

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
n/a

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Analysis of performance was based on 100% of the cases reported for this measure via claims for the PQRS program from 1/1/2014 to 12/31/2014. Data element validity and inter-rater reliability testing was performed on a random sample of this population (see section 1.5 and 2b.2.).

Performance score reliability testing was performed on 100% of cases reported for the PQRS program via claims and registry from 1/1/2013 to 12/31/2013.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The reporting of this measure is voluntary and total number of cases reported represents a small fraction of the total eligible population. Based on the 2014 PQRS Evaluation Report there were 26,978,892 eligible beneficiaries of which 2,212,704 (8.2%) were reported. The total number of eligible providers was 573,233 and 10.7% reported the measure.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The number of eligible providers reporting the measure is about 10.7% (3.6% in 2010, 4.5% in 2011, 1.8% in 2012, and 7.4% in 2013).

Because reporting is voluntary the reporting population cannot be said to be representative of the total eligible population. Generalizations to the overall eligible population should not be made.

Greater adoption of the measure, potentially via EHR reporting, will minimize potential bias caused by missing data from those who choose not to report.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

At the time of this submission, this measure is not currently being considered as eMeasure.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In an effort to reduce future variability in measure specification interpretation, the following changes will be reviewed:

1. Simplifying Numerator Quality codes [G8442 or G8939] from two G codes to one G code to identify the "Not Eligible" population.
2. Identify locations in the measure specification to emphasize documentation of the standardized tool

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

None

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Physician Quality Reporting System http://www.cms.gov/PQRS Payment Program Physician Quality Reporting System http://www.cms.gov/PQRS

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Public Use
Name: Physician Quality Reporting System (PQRS)
Sponsor: Centers for Medicare and Medicaid Services
Purpose and Geographical Area: PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries. Refer to the following link for additional information: http://www.cms.gov/PQRS

In 2014, there were 573,233 (10.7%) Eligible Professionals who could report NQF# 0420. In 2013, NQF #0420 was the 6th most reported measure within PQRS with 664,929 (7.4%) eligible professionals participating in reporting this measure.

Provider and Patients Statistics for program year 2014 (from "2014 Physician Quality Reporting System Program Monitoring and Evaluation Report"):
Providers
Eligible EPs in 2013-664,929
Eligible EPs in 2014=573,233

% of Eligible EPs who report in 2013=7.4%
% of Eligible EPs who report in 2014=10.7%

Beneficiaries
• Eligible Beneficiaries – 26,978,892
• Beneficiaries reported – 2,212,704
• % of Beneficiaries reported – 8.2%

Many types of providers/specialists report this measure as part of the PQRS as defined by the CPT codes in the measure specification.

Refer to section IV. Analysis of Claims Data in attached "NQF Endorsement Measurement Submission Summary Materials" document

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
n/a

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
n/a

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Progress on Improvement:
Average Performance Rates by Year based on data from "2014 Physician Quality Reporting System Program Monitoring and Evaluation Report":
2010 – 97.3% (3.6% of eligible providers)
2011 – 94.8% (4.5% of eligible providers)
2012 – 86.9% (1.8% of eligible providers)
2013 – 85.7% (7.4% of eligible providers)
2014 – 88.5% (10.7% of eligible providers)

Eligible Professionals by Year based on data from "2014 Physician Quality Reporting System Program Monitoring and Evaluation Report":
2010 – 170,678
2011 – 177,520
2012 – 705,787
2013 – 664,929
2014 – 573,233

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
It is difficult to say with certainty the reason for the decrease after 2010. These performance rates are submitted voluntarily by providers and cannot be generalized to the total population of eligible providers. The smaller group of early adopters may have been biased towards better performers. As a larger percentage of providers opt to report the measure we would expect to see the aggregate performance rate more closely estimate the true rate for the population.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0383 : Oncology: Plan of Care for Pain – Medical Oncology and Radiation Oncology (paired with 0384)
0676 : Percent of Residents Who Self-Report Moderate to Severe Pain (Short-Stay)
0677 : Percent of Residents Who Self-Report Moderate to Severe Pain (Long-Stay)
1628 : Patients with Advanced Cancer Screened for Pain at Outpatient Visits
1634 : Hospice and Palliative Care -- Pain Screening
1637 : Hospice and Palliative Care -- Pain Assessment

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

0050 : Osteoarthritis: Function and Pain Assessment/ National Committee for Quality Assurance
0306 : Back Pain: Patient Reassessment/ National Committee for Quality Assurance
0322 : Back Pain: Initial Visit/ National Committee for Quality Assurance
0341 : PICU Pain Assessment on Admission/ National Association of Children´s Hospitals and Related Institutions
0342 : PICU Periodic Pain Assessment/ National Association of Children´s Hospitals and Related Institutions
0523 : Pain Assessment Conducted/ Centers for Medicare and Medicaid Services
0675 : The Percentage of Residents on a Scheduled Pain Medication Regimen on Admission Who Self-Report a Decrease in Pain Intensity or Frequency (Short-stay)/ Centers for Medicare and Medicaid

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

Six related measures were identified that are not harmonized with NQF# 0420. The differences between these related measures and the submitted measure NQF# 0420 are listed below: 0383 - Oncology: Plan of Care for Pain – Medical Oncology and Radiation Oncology (paired with 0384 which is unrelated to and non-competing with 0420) - target population is specific to patients with a diagnosis of cancer currently receiving chemotherapy or radiation therapy who report having pain; 0383 does not include the use of a standardized pain assessment tool. Both measures are process measures. Both measures have outpatient care setting. 0676 - Percent of Residents Who Self-Report Moderate to Severe Pain (Short-Stay) – target population is specific to short - stay residents whereas 0420 has a broader outpatient population; 0420 is NOT a self-report measure, it is an eligible provider report; 0676 does not include the use of a standardized pain assessment tool; 0676 does not include documentation of a follow-up plan if pain is present; 0676 is an outcome measure whereas 0420 is a process measure. Care setting for 0676 is long term care/skilled nursing facilities whereas 0420 care setting is outpatient clinician office or outpatient rehabilitation. 0677 - Percent of Residents Who Self-Report Moderate to Severe Pain (Long-Stay) – target population is specific to long - stay residents whereas 0420 has a broader outpatient population; 0420 is NOT a self-report measure, it is an eligible provider report; 0677 does not include the use of a standardized pain assessment tool; 0677 does not include documentation of a follow-up plan if pain is present; 0677 is an outcome measure whereas 0420 is a process measure. Care setting for 0677 is long term care/skilled nursing facilities whereas 0420 care setting is outpatient clinician office or outpatient rehabilitation. 1628 - Patients with Advanced Cancer Screened for Pain

at Outpatient Visits - target population is specific to patients with a diagnosis of advanced cancer; 1628 does not include a follow-up plan if pain is present; Both 1628 and 0420 are process measures; Both measures have outpatient care setting. 1634 - Hospice and Palliative Care -- Pain Screening: target population has no age parameters whereas 0420 has an age range (> 18 yrs.); 1634 target population is specific to hospice and palliative care patients whereas 0420 is not diagnosis specific; 1634 does not include documentation of a follow-up plan if pain is present; Both 1634 and 0420 are process measures; Care setting for 1634 is restricted to Hospice/Hospital/Acute Care Facility, whereas 0420 care setting is outpatient clinician office or outpatient rehabilitation. 1637 – Hospice and Palliative Care—Pain Assessment- target population has no age parameters whereas 0420 has an age range (> 18 yrs.); 1637 target population is specific to hospice and palliative care patients whereas 0420 is not diagnosis specific; 1637 measure focus is clinical assessment within 24hrs of positive screening for pain; 0420 measure focus is performing a screening and a documented follow-up plan not just limited to a clinical assessment; Both are process measures; Care setting for 1637 is restricted to Hospice/Hospital/Acute Care Facility; whereas 0420 care setting is outpatient clinician office or outpatient rehabilitation.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
There are no competing measures.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** NQF_Endorsement_Measurement_Submission_Summary_Materials.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Sophia, Autrey, Sophia.autrey@cms.hhs.gov, 410-786-1158-
Co.3 **Measure Developer if different from Measure Steward:** Centers for Medicare & Medicaid Services
Co.4 **Point of Contact:** Sophia, Autrey, Sophia.autrey@cms.hhs.gov, 410-786-1158-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Through a collaborative process, the Technical Expert Panel (TEP) reviewed the current 2016 measure specifications (description, numerator, denominator, definitions, clinical recommendation, and environmental scan).

Camielle Call, LCSW, MSW, Social Worker, University of Alaska Southeast

Jean Carter, PhD, Psychologist, Washington Psychological Center, P.C.

Ann Marie Feretti, Adv, MS, OTR/L, CHT, Occupational Therapist, PROACTIVE Physical & Hand Therapy

Craig S. Little, DC, FACO, Chiropractor, Independent Practice

Elisa Marks, OTR/L, CHT, Occupational Therapist, Center for Health Enhancement and Rehabilitation (CHEAR)

Gregory M. Martino, PhD, Clinical Psychologist, Independent Practice

William Glancey, Patient/Caregiver representative

Christine Goertz, DC. PhD, Chiropractor, Vice Chancellor for Research and Health Policy, Palmer College of Chiropractic

Deepthi Saxena, MD, Physiatrist, Medical Director, Affiliated Medical Rehabilitation

Donna M. Ulteig, LCSW, Licensed Clinical Social Worker, Psychiatric Services, SC

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2008

Ad.3 **Month and Year of most recent revision:** 09, 2015

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 09, 2016

Ad.8 **Additional Information/Comments:**

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2614
De.2. **Measure Title:** CoreQ: Short Stay Discharge Measure
Co.1.1. **Measure Steward:** American Health Care Association
De.3. **Brief Description of Measure:** The measure calculates the percentage of individuals discharged in a six month time period from a SNF, within 100 days of admission, who are satisfied (see: S.5 for details of the time-frame). This patient reported outcome measure is based on the CoreQ: Short Stay Discharge questionnaire that utilizes four items.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:
(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.
The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007). Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.
The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming,

one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider. As part of the U.S. Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Short Stay Discharge questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007). This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Short Stay Discharge questionnaire measure is needed to improve the care for short stay SNF patients.

Furthermore, improving the care for short stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Short Stay Discharge measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Short Stay Discharge measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.
CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf
CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.
Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.
Institute of Medicine (2001). Improving the Quality of Long Term Care, National Academy Press, Washington, D.C., 2001.
Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).
MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.
Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.
Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

S.4. **Numerator Statement:** The measure assesses the number of patients who are discharged from a SNF, within 100 days of admission, who are satisfied. The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the four questions on the CoreQ: Short Stay Discharge questionnaire.

S.7. **Denominator Statement:** The denominator includes all of the patients that are admitted to the SNF, regardless of payor source, for post-acute care, that are discharged within 100 days; who receive the survey (e.g. people meeting exclusions do not receive a questionnaire) and who respond to the CoreQ: Short Stay Discharge questionnaire within the time window (See: S.5).

S.10. **Denominator Exclusions:** Exclusions used are made at the time of sample selection and include:

(1) Patients who died during their SNF stay;

(2) Patients discharged to a hospital, another SNF, psychiatric facility, inpatient rehabilitation facility or long term care hospital;

(3) Patients with court appointed legal guardian for all decisions;

(4) Patients discharged on hospice;

(5) Patients who left the nursing facility against medical advice (AMA);

(6) Patients who have dementia impairing their ability to answer the questionnaire defined as having a BIMS score on the MDS as 7 or lower. [Note: we understand that some SNCCs may not have information on cognitive function available to help with sample selection. In that case, we suggest administering the

survey to all residents and assume that those with cognitive impairment will not complete the survey or have someone else complete on their behalf which in either case will exclude them from the analysis.] (7) Patients who responded after the two month response period; and (8) Patients whose responses were filled out by someone else.
Measure Type: PRO **Data Source:** Healthcare Provider Survey **Level of Analysis:** Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report
1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.
 	
Summary of evidence:
 	
• This is a patient-reported outcome measure of patient satisfaction. The developer provides a diagram and a table demonstrating the links between structures and/or processes and the outcomes that have been found to influence patient satisfaction, and the final patient reported outcome of satisfaction.
• The developer notes that "Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management"
• The developer states "We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care" and notes that measuring patient satisfaction is required for person-centered care for three reasons:
  ○ Measuring satisfaction is necessary to understand patient preferences.
  ○ Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
  ○ Satisfaction information can help facilities improve the quality of care they provide
 	
Guidance from the Evidence Algorithm
PRO-based measure (Box 1) → Relationship between the outcome and at least one healthcare action is identified and supported by the rationale (Box 2) → PASS
 	
Question for the Committee:
▪ *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**
1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provided the following information on performance gap:

- Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes.
- Studies show a link between patient satisfaction and the following health-related behaviors:
    - Keeping follow-up appointments
    - Disenrollment from health plans
    - Litigation against providers

The developer provided performance scores based on 10,319 responses from 265 facilities that met the inclusion criteria (20 valid responses and 30% response rate). The scores include tables by age and gender.

Facility Level Performance Distribution

Questionnaire Item	Observation	Mean	Standard Deviation	Minimum	Maximum
1. In recommending this facility to your friends and family, how would you rate it overall?	265	3.61	.44	1.8	4.6
2. Overall, how would you rate the staff?	265	3.80	.38	2	5
3. How would you rate the care you receive?	265	3.68	.43	1.8	5
4. How would you rate how well your discharge needs were met?	265	3.65	.43	2	5

Overall Descriptive Information for the CoreQ: Short Stay Discharge Measure

	min	p25	p50	p75	max
Summary Score	25.0	75.0	82.5	88.6	100.0

Disparities

The developer says differences in scores based on SDS categories were not statistically significant:

- By race/ethnicity, whites averaged a score of 83.3, Blacks or African-Americans averaged a score of 83.4, and Asians 83.4
- By highest education level those with those high school but who did not graduate averaged 83.2, high school graduates averaged 83.1, those with some college or a 2-year degree averaged 82.9, 4 year college graduates averaged 83.1, and those with more than 4 year college degree averaged 83.8
- By age group, residents younger than 65 years old averaged 70.0, those 65-74 averaged 84.8, those 75-84 averaged 84.6, and those older than 85 averaged 87.1
- by gender, males averaged a score of 89.2 and females averaged a score of 81.4

However, research over the last 20 years has consistently found poorer care in facilities with high minority populations and that nursing homes remain segregated, with black patients concentrated in poorer-quality homes (as measured by staffing ratios, performance, and are more financially vulnerable).

The developer did not risk adjust this measure for SES because “adjusting for racial status has the unintended effect of

adjusting for poor quality providers not to differences due to racial status and not within-provider discrimination." They further comment: "…lower satisfaction scores for both Caucasian and Blacks and other ethnicities are likely to increase as the proportion of black residents increases in a SNF, indicating that the best measure of racial disparities in satisfaction rates is one that measures scores at the facility level. That is, ethnic and social economic status differences are related to inter-facility differences not to intra-facility differences in care. Therefore, the literature suggests that racial status should not be risk adjusted otherwise one is adjusting for the poor quality of the SNFs rather than differences due to racial status."

Meaningfulness to the Target Population (PRO-PM):

- The developer provided an overview Specific to the CoreQ: Short Stay Discharge questionnaire, the importance of the satisfaction areas assessed were examined with focus groups of residents and family members. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. Table 1c.5 in the appendix shows the score of the importance for question included in the CoreQ: Short Stay Discharge questionnaire. The overall ranking used was 10=Most important and 1=Least important. The final four questions included in the measure had average scores ranging from 9.35 to 9.69; this clearly shows that the respondents value the items used in the CoreQ: Short Stay Discharge measure.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**Evidence provided and this is an area where having a standard survey would help fill a call for more patient experience assessment in skilled nursing facilities.

**Conceptual framework outlining types of process measures that might influence patient satisfaction with SNF provided. A number of studies are cited to support conceptual framework, linking satisfaction with responsiveness of management, staff competence, staffing levels, and care/concern of staff.

1b. Performance Gap

Comments:

**The data shows opportunity. It was also reported by race, education, age and gender. It was not adjusted for SES.

**Developers cite a literature review from 2007 indicating average patient satisfaction scores in long term care facilities ranging from 60-80%. They also note substantial variation in instruments used to measure satisfaction limiting generalizability/synthesis of findings.

In a study of 282 nursing facilities (n patients=10,319, response rate=~30%), facility level scores were slightly favorably skewed but showed adequate distribution to indicate a performance gap (interquartile range = 75.6-88.6).

Disparities were evident by age (older respondents were satisfied) and gender (males more satisfied) but not for race/ethnicity or education.

1c. Pro-PM

Comments:
**Focus groups of patients and families were included in the development of the survey.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): CoreQ: Short Stay Discharge questionnaire – health care provider

Specifications:

- The level of analysis is facility.
- The measure result is a non-weighted percentage score:
  - The numerator is the number of patients who are discharged from a SNF, within 100 days of admission, who are satisfied.
  - The denominator is all patients that are admitted to the SNF, regardless of payor source, for post-acute care, that are discharged within 100 days; who receive the survey (e.g. people meeting exclusions do not receive a questionnaire) and who respond to the CoreQ: Short Stay Discharge questionnaire within the time window.
  - "Satisfied" individuals are those that have an average satisfaction score of =>3 for the four questions on the CoreQ: Short Stay Discharge questionnaire
- There is no data dictionary.
- A calculation algorithm is described.
- The measure is not risk adjusted or stratified.
- There are 8 exclusions:
  - Patients who died during their SNF stay;
  - Patients discharged to a hospital, another SNF, psychiatric facility, inpatient rehabilitation facility or long term care hospital;
  - Patients with court appointed legal guardian for all decisions;
  - Patients discharged on hospice;
  - Patients who left the nursing facility against medical advice (AMA);
  - Patients who have dementia impairing their ability to answer the questionnaire defined as having a BIMS score on the MDS as 7 or lower.
  - Patients who responded after the two month response period; and
  - Patients whose responses were filled out by someone else.
- The calculation of exclusion criteria is specified and includes MDS and nursing home facility health information system data.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data elements were tested using a test-retest methodology. The survey was sent out and responses received from 853 patients; 100 were re-surveyed one month later. The distribution of responses and the correlation between the original and follow-up scores were then calculated.
- Person/questionnaire level was tested using the same test-retest methodology.
- The stability of the facility-level score was tested using bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score.

Results of reliability testing

Results for each level of testing are presented.

- Data element testing showed very high levels of agreement and no statistically significant difference in the responses to each question between the original and re-test results.

Average Percent Agreement between 1st and 2nd Administered Surveys

Questionnaire Item	Percent Agreement
1. In recommending this facility to your friends and family, how would you rate it overall?	96.8%
2. Overall, how would you rate the staff?	97.8%
3. How would you rate the care you receive?	98.2%
4. How would you rate the discharge process?	98.2%

- Person/questionnaire level agreement showed very high levels of agreement and no statistically significant difference in the responses to each question

		Re- administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
	Poor (1) or Average (2)	98.5%	98%

Pilot Response	Good (3), Very Good (4), or Excellent (5)	98.5%	99%

Measure level testing also demonstrated agreement:

- 17.82% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample
- 38.14% were within 3 percentage points
- 61.05% were within 5 percentage points
- 87.05% were within 10 percentage points

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empiric testing- yes (box 2) -> with measure score – yes (box 4) – appropriate method – yes (box 5) – Level of certainty or confidence in the performance measure scores (box 6): HIGH

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

1. Validity testing of the questionnaire format used in the CoreQ: Short Stay Discharge questionnaire
    - Face validity evaluated via literature review and review of 12 commonly used satisfaction surveys; also

examined face validity of domains and the response scale, using 40 patients in 5 nursing homes. The Flesch-Kinkaid scale was used to determine if patients understood the questions.

2. Testing the items for the CoreQ: Short Stay Discharge questionnaire;

- Exploratory factor analysis (EFA) were used to further refine the pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items.

3. To determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Short Stay Discharge measure);

- Correlation analysis and a factor analysis conducted on items

4. Validity Testing for the CoreQ: Short Stay discharge measure.

- Developers examined correlation between the four items in the measure and all of the items on the pilot instrument.
- Also examined correlations between the CoreQ: Short Stay Discharge measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, ii) several other quality metrics from Nursing Home Compare, iii) risk adjusted discharge to community measure and iv) risk adjusted PointRight® Pro 30™ Rehospitalizations

Validity testing results:

Results for each level of validity testing are provided. The developer interpretation of results is as follows:

1. Validity Testing for the Questionnaire Format used in the CoreQ: Short Stay Discharge Questionnaire
   A. The literature review shows that domains used in the Pilot CoreQ: Short Stay Discharge questionnaire items have a high degree of both face validity and content validity.
   B. Patients overall rankings, show the general “domain” areas used indicates a high degree of both face validity and content validity.
   C. The results show that 100% of residents are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.
   D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the items.
2. Testing the Items for the CoreQ: Short Stay Discharge Questionnaire
   A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks.
   B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only “concept” being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of “customer satisfaction”. This testing indicates a high degree of criterion validity.
3. Determine if a Sub-Set of Items Could Reliably be Used to Produce an Overall Indicator of Satisfaction (The Core Q: Short Stay Discharge Measure).
   A. Using the correlation information of the Core Q: Short Stay Discharge questionnaire (22 items) and the 4 items representing the CoreQ: Short Stay Discharge questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.
   B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only “concept” being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ

instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Short Stay Discharge Measure.

A. The correlation of the 4 item CoreQ: Short Stay Discharge measure summary score (identified elsewhere in this document) with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.94.

That is, the correlation score between actual the "CoreQ: Short Stay Discharge Measure" and all of the 22 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 4 items or the 22 item Pilot questions.

This indicates that the CoreQ: Short Stay Discharge instrument summary score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B.

(i) Relationship with CASPER Quality Indicators

The 8 CASPER quality indicators had a low to moderate level of negative correlation with the CoreQ: Short Stay Discharge measure. Those that correlate have a clear conceptual link with short stay, and those that do not are more associated with long stay residents or have unclear conceptual links to short stay customer satisfaction. The CASPER quality indicators that correlate with the CoreQ Short Stay Discharge score are any deficiency citations (-0.11; p=0.07), pressure ulcers (-0.22, p<0.01) and antidepressants (+0.13, p=0.03); those that do not correlate are physical restraints (-0.01, p=0.91), catheterization (-0.04, p=0.56), antipsychotic medications (-0.06, p=0.32), antianxiety medications (0.08, p=0.19), and hypnotic medications (0.04, p=0.46). This testing indicates a moderate degree of construct validity and convergent validity.

(ii) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings and staffing levels

The Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels all had a moderately high levels of correlation and in the direction predicted with the CoreQ: Short-Stay Discharge measure. These correlations range from ± 0.120 to 0.330. The CoreQ: Short-Stay Discharge measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.330 for RN hours per resident-day and 0.305 for total staffing hours per resident day. This testing indicates a high degree of construct validity and convergent validity.

(iii) Relationship with the risk-adjusted Discharge to Community Measure

The risk-adjusted Discharge to community measure was negatively correlated to the CoreQ: Short Stay Discharge measure. The correlations were small ranging from -0.05 to -0.16. This was not as hypothesized which may be related to some SNFs that specialize in long stay, have very low discharge to community rates as admissions do not have a plan to go home.

(iv) Relationship with the risk adjusted PointRight® Pro 30™ Rehospitalizations

The risk-adjusted PointRight® Pro 30™ Rehospitalizations was negatively correlated to the CoreQ: Short Stay Discharge measure. The correlations were modest ranging from -0.22 to -0.31, and all of them were statistically significant at the p-value of 0.05. This is expected because lower rehospitalization rates (an indicator of high quality) are associated with higher satisfaction. This was as hypothesized. This testing indicates a reasonable degree of construct validity and convergent validity.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- An expert panel advised the developer on exclusions. They were advised to exclude patients who died, patients who were discharge to a hospital, patients with durable power of attorney for all decisions, patients on hospice, patients with low BIMS scores, and patients who left against medical advice, which the developer reports are all standard exclusions for satisfaction surveys.
- The first analysis included responses from 10,319 patients. Exclusions were tracked and the following reported:
  - 1,970 patients (19.1%) discharged to the hospital;
  - 5 (0.05%) discharged to hospice; and,
  - 10 (0.09%) expired.
  - Patients that had left against medical advice or had a durable power of attorney were not tracked in this sample.
- The second analysis included 100 nursing homes and data from the first 1000 patients.
  - 791 patients (7.9%) were discharged to the hospital;
  - 48 (0.48%) were discharged to hospice;
  - 41 (0.41%) expired;
  - 23 (0.23%) left against medical advice; and
  - 46 (0.46%) had a durable power of attorney.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Risk adjustment summary

The developers provide the following rationale for no risk-adjustment:
"No research (to date) has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS® Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical or theoretical or empirical risk adjusted or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to resident characteristics and the satisfaction scores."

1RTI International, Harvard University, RAND Corporation. CAHPS Instrument for Persons Residing in Nursing Homes, Final Report to CMS, CMS Contract No. CMS-01-01176, Sept. 2003.

Questions for the Committee:

- *A justification for no risk adjustment is provided. Is there any evidence that contradicts the developer's rationale and analysis?*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states "The CoreQ Short Stay Discharge scores reflect practical and meaningful differences in quality between facilities. The histogram in Section 2b5.2 (figure 1b.2) shows that the distribution of summary scores is quite wide, indicating the scores can be used to differentiate facilities of varying levels of customer satisfaction quality."

- Of the 265 facilities in the test population, scores ranged from 1 facility scoring 15-20% to 11 facilities scoring 90-95%.

Question for the Committee:
Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer states missing data was uncommon (<13% of any one of the 4 items). For patients with one missing data point (from the 4 items included in the CoreQ: Short Stay Discharge questionnaire) imputation is utilized (representing the average value from the other available data points). Patients with more than one missing data point, are excluded.

Preliminary rating for validity: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Guidance from the Reliability Algorithm
Specifications consistent with evidence (Box 1): Yes →Potential threats to validity assessed (Box 2): Yes →Empirical validity testing performed using measure as specified (Box 3): Yes →Validity testing with computed performance measure score (Box 6): Yes →Method Described appropriate (box 7): Yes →Level of certainty or confidence that the performance measure score is a valid indicator of quality (Box 8): High

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 Specifications
Comments:
**The questions used are correlated with overall satisfaction.
**2a1. Data elements clearly defined; person and facility level scoring provided.

Exclusions may limit generalizability of satisfaction results to a small proportion of nursing facility residents. Consistency of implementation of instrument across facilities may be compromised by low response rates.

2a.2 Reliability Testing
Comments:
**Results were retested with similar results.
**Developers appear to perform test-retest reliability but give 1 month as the testing interval and give % agreement but not intra-rater reliability at the patient level, or ICCs at the facility level.

The facility-level bootstrapping procedure cannot be adequately interpreted without a pre-specifying minimally meaningful difference. Further, the variation in patient sample size by facility (npatients=20-196) will substantially alter the between facility reliability estimates.

Cronbach's alpha provided in Table 2b2.3.g. suggests high internal consistency reliability for the items chosen but developers do not provide the item-to-total correlation coefficients for all 22 items to support their item choices.

2b2. Validity Testing

Comments:
**Patient level - The shorter survey was tested against a longer survey with a high degree of correlation.
**Face validity was used to generate domains of observables from 12 "commonly used" satisfaction instruments and to evaluate measure content.

Patients (n=40) were also used to rank 22 candidate items for inclusion in the final instrument from most to least important. Cognitive testing was used to confirm respondents' understanding of item content and response options. Pilot testing on a convenience sample (n=853) support choice of response options. Factor analysis appear to support a single dimension, although varimax or oblique rotated factor results were noted as performed but not provided. The decision to eliminate items cannot be effectively evaluated, nor can potential multi-dimensionality of the construct. That the reliability coefficient(?) appeared to be stable for 4 vs. 22 items is odd since the calculation of that coefficient includes K-of items and increases with the inclusion of more items (a reliability not validity issue).

Construct validation using CASPER quality indicators, Nursing Home Compare, and risk-adjusted Discharge to Community (NQF #2858) suggests no or minimal relationship between those measure and satisfaction.

A significant negative relationship between satisfaction and risk adjusted readmissions [NQF #2375] do appear to provide some evidence for construct validity (greater satisfaction with lower rehospitalizations). However, one item ("How would you rate how well your discharge needs were met?") appears to be in the opposite direction, unless there is a typo, this is confusing and no interpretation is provided by the developer.

2b3.-2b7. Validity Testing
Comments:
**Exclusions clearly stated. One consideration would be to consider whether another person filling the survey out would be an exclusion - if the goal is patient and family centered care and the patient is unable to fill out the survey, the perceptions by the family could be important.
**2b3. Exclusions did not appear to represent a problem in compromising the application or generalizability of the measure.

2b4. No risk adjustment is to be done despite evidence for age, gender differences noted in 1b.

2b5. Meaningful differences should have been provided as facility-level effect sizes and were not.

2b7. Low response rates represent a potentially serious compromise in data quality and completeness.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The collection instrument is the CoreQ: Short Stay Discharge questionnaire and Resident Assessment Instrument Minimum Data Set (MDS) version 3.0.
- This is a patient satisfaction survey conducted via mailed survey.
- No fees required to use the measure; the developer did not indicate if there are fees associated with the use of the survey.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

○ *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**Response rates weren't provided but the shorter survey might lend itself to a higher response rate.

It also wasn't stated as to whether the survey is available in other languages.

**Although a short questionnaire, the low response rate may seriously limit feasibility and usability.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

- AHCA Quality Initiative: https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx
- Massachusetts Senior Care (150 facilities)
- Satisfaction Vendors (10 national companies)

Quality Improvement (Internal to the specific organization)

- Large Nursing Home Chain

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**

OR

Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

Not in use for accountability program, but ACHA plans to begin public reporting of the CoreQ measures as part of their Quality Initiative 2016-2018 (9,600 SNFs)

Improvement results [Impact/trends over time/improvement]

Unexpected findings (positive or negative) during implementation
None reported

Potential harms
The developer states, "There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make those patients further dissatisfied."

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use
Comments:

**It is not being publicly reported. It could lend itself to that in the future.

Criterion 5: Related and Competing Measures

Related or competing measures
The developers cited potential relatedness/competing with the CAHPS Nursing Home surveys, however; the measures derived from Nursing Home CAHPS have recently lost endorsement. AHRQ has communicated lack of resources to maintain the measures, and they are not currently in use in any federal program.

Harmonization
N/A

-

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A

Measure Title: *CoreQ: Short Stay Discharge Measure*

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: **Instructions**

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence sub criterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☒Patient-reported outcome (PRO): Customer Satisfaction

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

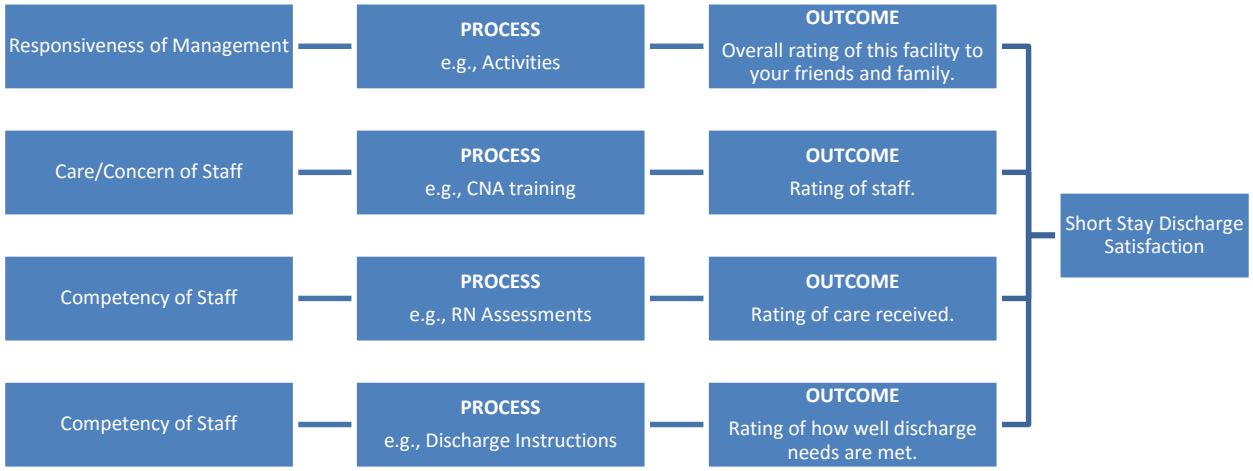
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Short stay discharge satisfaction can be looked at as the outcome for a number of structures and processes within skilled nursing care centers. Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management (National Research Corporation, 2014).



Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the Health Professions,* 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

National Research Corporation. (2014). 2014 National Research Report Empowering Customer-Centric Healthcare Across the Continuum.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

In a review of the satisfaction literature, Castle (2007) noted that the structure, process, outcome model was most commonly used to identify the factors that influence satisfaction. The table below provides the structure and process drivers that are associated with our stated outcome of customer satisfaction.

Table 1a.2.1: The structure and process drivers associated with short stay discharge satisfaction.

Authors	***Structure or Process and Driver of Short Stay Discharge Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Reinhardt, et al., 2014	Process Responsiveness of management and care/concern of staff	Conversations regarding end-of-life care options with family members show higher overall satisfaction with care and more use of advance directives.	Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. *Journal Of Social Work In End-Of-Life & Palliative Care*. 10(2):112-26.
Lin et al., 2014.	Process Competency of Staff	Significant difference for overall resident satisfaction with higher perceived service quality.	Lin, J., Hsiao, C.T., Glen, R., Pai, J.Y., & Zeng, S.H. (2014). Perceived service quality, perceived value, overall satisfaction and happiness of outlook for long-term care institution residents. *Health Expectations*. 17(3):311-20.
Van Uden et al. (2013).	Process Competency of Staff	For nursing home residents with dementia improved symptom management is associated with higher satisfaction with care.	Van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. *International Psychogeriatrics.* 25(10):1697-707.
Li et al. (2013).	Structure Competency of	Higher overall nursing home satisfaction scores were associated with higher nursing	Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care

	Staff	staffing levels and fewer deficiency citations.	was generally high during 2005-09, with some variability across facilities. *Health Affairs*. 32(8):1416-25.
Authors	***Structure or Process***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Brownie & Nancarrow (2013).	Structure & Process Responsiveness of management and Care/concern of staff	Implementation of person-centered care is associated with higher levels of satisfaction.	Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. *Clinical Interventions In Aging*. 8:1-10.
Kleijer et al., 2014	Process Competency of staff	Residents perceive a low level of quality of care in centers where there is a high level of antipsychotic use.	Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts, A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.
Bishop et al., 2008	Structure Care/concern of staff	CNA's that receive a good supervision are more committed to staying in their jobs. This commitment in turn leads to positive relationships with resident and higher resident satisfaction.	Bishop, C., Weinberg, D., Leutz, W., Dossa, A., Pfefferle, S., & Zincavage, R. (2008). Nursing assistants' job commitment: Effect of nursing home organizational factors and impact on resident well-being. *The Gerontologist*, 48(1), 36-45.

Kayser-Jones et al., 1999	Structure Responsiveness of management and competency of staff	Higher levels of RN and LPN staffing have been associated with better quality outcomes such as ADL maintenance and hydration. Centers that have a family council in addition to the required resident council have higher resident satisfaction.	Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision. *Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the Health Professions,* 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts, A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.

Bishop, C., Weinberg, D., Leutz, W., Dossa, A., Pfefferle, S., & Zincavage, R. (2008). Nursing assistants' job commitment: Effect of nursing home organizational factors and impact on resident well-being. *The Gerontologist*, 48(1), 36-45.

Lucas, J.A., Lowe, T.J., Robertson, B., Akincigil, A., Sambamoorthi, Q., Bilder, S., Paek, E.K., & Crystal, S. (2007). The relationship between organizational factors and resident satisfaction with nursing home care and life. *Journal of Aging & Social Policy,* 19(2), 125-151.

Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision. *Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Kane, R.L., & Kane, R.A. (2001). What older people want from long-term care, and how can they get it. *Health Affairs*, 20(6), 114-127.

Westat. Resident experience with nursing home care: A literature review.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 2614
De.2. **Measure Title:** CoreQ: Short Stay Discharge Measure
Co.1.1. **Measure Steward:** American Health Care Association
De.3. **Brief Description of Measure:** The measure calculates the percentage of individuals discharged in a six month time period from a SNF, within 100 days of admission, who are satisfied (see: S.5 for details of the time-frame). This patient reported outcome measure is based on the CoreQ: Short Stay Discharge questionnaire that utilizes four items.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:
(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.
The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.
The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.
As part of the U.S. Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Short Stay Discharge questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over

the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007). This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Short Stay Discharge questionnaire measure is needed to improve the care for short stay SNF patients.

Furthermore, improving the care for short stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Short Stay Discharge measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Short Stay Discharge measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.
CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf
CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.
Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.
Institute of Medicine (2001). Improving the Quality of Long Term Care, National Academy Press, Washington, D.C., 2001.
Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).
MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.
Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.
Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

S.4. **Numerator Statement:** The measure assesses the number of patients who are discharged from a SNF, within 100 days of admission, who are satisfied. The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the four questions on the CoreQ: Short Stay Discharge questionnaire.
S.7. **Denominator Statement:** The denominator includes all of the patients that are admitted to the SNF, regardless of payor source, for post-acute care, that are discharged within 100 days; who receive the survey (e.g. people meeting exclusions do not receive a questionnaire) and who respond to the CoreQ: Short Stay Discharge questionnaire within the time window (See: S.5).
S.10. **Denominator Exclusions:** Exclusions used are made at the time of sample selection and include:

(1) Patients who died during their SNF stay;

(2) Patients discharged to a hospital, another SNF, psychiatric facility, inpatient rehabilitation facility or long term care hospital;

(3) Patients with court appointed legal guardian for all decisions;

(4) Patients discharged on hospice;

(5) Patients who left the nursing facility against medical advice (AMA);

(6) Patients who have dementia impairing their ability to answer the questionnaire defined as having a BIMS score on the MDS as 7 or lower. [Note: we understand that some SNCCs may not have information on cognitive function available to help with sample selection. In that case, we suggest administering the survey to all residents and assume that those with cognitive impairment will not complete the survey or have someone else complete on their behalf which in either case will exclude them from the analysis.]

(7) Patients who responded after the two month response period; and

(8) Patients whose responses were filled out by someone else.

De.1. **Measure Type:** PRO
S.23. **Data Source:** Healthcare Provider Survey
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not Applicable

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
CoreQ_Short_Stay_Evidence_Final-635949676534319959.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.

(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.

(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).

Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.

The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.

As part of the U.S. Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Short Stay Discharge questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007). This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Short Stay

Discharge questionnaire measure is needed to improve the care for short stay SNF patients.

Furthermore, improving the care for short stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Short Stay Discharge measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Short Stay Discharge measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.
CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf
CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.
Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.
Institute of Medicine (2001). Improving the Quality of Long Term Care, National Academy Press, Washington, D.C., 2001.
Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).
MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.
Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.
Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the*

subcriterion on improvement (4b.1) under Usability and Use.
The appendix (section 1b.2) provides data sourced from 282 nursing facilities that are part of one large chain and include responses from 10,319 patients. The data were collected from June 2014 through September 2014.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We did not risk adjust the measure by sociodemographic status due to no statistically significant differences (at the 5% level) in the scores between the SDS categories. See Table 2b4.4b.b in the Testing section. By race, whites averaged a score of 83.3, Blacks or African-Americans averaged a score of 83.4, and Asians 83.4; there were no observations for Native Hawaiians or other Pacific Islanders, American Indian or Alaskan Natives (Table 2b4.4b.c in the Testing section). By highest education level those with those high school but who did not graduate averaged 83.2, high school graduates averaged 83.1, those with some college or a 2-year degree averaged 82.9, 4 year college graduates averaged 83.1, and those with more than 4 year college degree averaged 83.8 (Table 2b4.4b.c in the Testing section). By age group, residents younger than 65 years old averaged 70.0, those 65-74 averaged 84.8, those 75-84 averaged 84.6, and those older than 85 averaged 87.1 (Table 1b.4.a in the Appendix). Furthermore, by gender, males averaged a score of 89.2 and females averaged a score of 81.4 (Table 1b.4.b in the Appendix).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Multiple studies in the past twenty years have examined racial disparities in the care of nursing facility residents and have consistently found poorer care in facilities with high minority populations (Fennell et al., 2000; Mor et al., 2004; Smith et al., 2007). Work on racial disparities in nursing facilities' quality of care between elderly white and black residents within nursing facility has shown clearly that nursing homes remain relatively segregated and that specifically nursing home care can be described as a tiered system in which blacks are concentrated in marginal-quality homes (Li, Ye, Glance & Temkin-Greener, 2014; Fennell, Feng, Clark & Mor, 2010; Li, Yin, Cai, Temkin-Greener, Mukamel, 2011; Chisholm, Weech-Maldonado, Laberge, Lin, & Hyer, 2013; Mor et al., 2004; Smith et al., 2007). Such homes tend to have serious deficiencies in staffing ratios, performance, and are more financially vulnerable (Smith et al, 2007; Chisholm et al., 2013). Based on a review of the nursing facility disparities literature, Konetzka and Werner concluded that disparities in care are likely related to this racial and socioeconomic segregation as opposed to within-provider discrimination (Konetzka and Werner 2009). This conclusion is supported, for example, by Grunier and colleagues who found that as the proportion of black residents in the nursing home increased the risk of hospitalization among all residents, regardless of race, also increased (Grunier et al., 2008). Thus, adjusting for racial status has the unintended effect of adjusting for poor quality providers not to differences due to racial status and not within-provider discrimination.
Therefore, lower satisfaction scores for both Caucasian and Blacks and other ethnicities are likely to increase as the proportion of black residents increases in a SNF, indicating that the best measure of racial disparities in satisfaction rates is one that measures scores at the facility level. That is, ethnic and social economic status differences are related to inter-facility differences not to intra-facility differences in care. Therefore, the literature suggests that racial status should not be risk adjusted otherwise one is adjusting for the poor quality of the SNFs rather than differences due to racial status.

Chisholm L, Weech-Maldonado R, Laberge A, Lin FC, Hyer K. (2013). Nursing home quality and financial performance: does the racial composition of residents matter? Health Serv Res;48(6 Pt 1):2060–2080.

Fennell ML, Feng Z, Clark MA, Mor V. (2010). Elderly Hispanics more likely to reside in poor-quality nursing homes. Health Aff (Millwood);29(1):65–73.

Grabowski, D.C. (2004). The admission of Blacks to high-deficiency nursing homes. Medical Care 42(5): 456-464.

Gruneir, A., Miller, S. C., Feng, Z., Intrator, O., & Mor, V. (2008). Relationship between state Medicaid policies, nursing home racial composition, and the risk of hospitalization for black and white residents. Health Services Research, 43(3), 869-881.

Konetzka, R. T., & Werner, R. M. (2009). Review: Disparities in long-term care building equity into market-based reforms. Medical Care Research and Review, 66(5), 491-521.

Li Y, Yin J, Cai X, Temkin-Greener J, Mukamel DB. (2011). Association of race and sites of care with pressure ulcers in high-risk nursing home residents. JAMA;306(2):179–186.

Li Y, Ye Zhiqiu, Glance, Laurent & Temkin-Greener, Helena. (2014). Trends in family rating experience with care and racial disparities among Maryland nursing homes. Med Care, 52(7): 641-648.

Mor, V., Zinn, J., Angelelli, J., Teno, J. M., & Miller, S. C. (2004). Driven to tiers: socioeconomic and racial disparities in the quality of nursing home care. Milbank Quarterly, 82(2), 227-256.

Smith, D. B., Feng, Z., Fennell, M. L., Zinn, J. S., & Mor, V. (2007). Separate and unequal: racial segregation and disparities in quality across US nursing homes. Health Affairs, 26(5): 1448-1458.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, Patient/societal consequences of poor quality

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The definition of quality in a nursing facility has shifted from a focus on structure and process criteria to clinical outcomes, resident satisfaction, and quality of life. This shift was first supported by nursing home reform legislation included in the Omnibus Budget Reconciliation Act of 1987 (OBRA, 1987). Furthering the movement, the Institute of Medicine (IOM) put the patient as central to the care system (Castle, 2007; IOM, 2001) – necessitating the collection of satisfaction information. As mentioned previously (see 1b.1), a focus on person-centered care and satisfaction is also evident in the Quality Assurance & Performance Improvement Program (QAPI) for nursing facilities and proposed Reform Requirements for Long-Term Care Facilities (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015).

Measuring and reporting satisfaction of nursing home care is important in many ways. First, residents are more likely to follow medical advice when they rate their care as satisfactory (Hall, Milburn, Roter, & Daltroy, 1998). Second, because resident satisfaction can influence the quality of care provided and the outcomes of treatment

(Hudak and Wright 2000), satisfaction surveys can be used as measures of clinical and organizational accountability. Third, measuring and reporting resident satisfaction can help nursing facilities identify and improve aspects of quality. Furthermore, if publicly released, information on satisfaction with care can help elders and their families choose a nursing facility.

Several research efforts have concluded consumer satisfaction is an important indicator of quality of care in nursing homes (Gesell, 2001; Bangerter et al. 2016; Shippee et al 2015; Kajonius and Kazemi, 2016). In addition, other studies have concluded nursing resident satisfaction data provides information about quality of care that is different from clinician perspectives and clinical indicators (Berlowitz, Du, Kazis, & Lewis, 1993; Riccio 2000; Uman & Urman, 1997). This exemplifies the need for resident satisfaction data to achieve person-centered care. Only by hearing from the patient can we ensure the care provided is person-centered.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Bangerter, L.R., Heid, A.R., Abbott, K, & Van Haitsma, K. (2016). Honoring the Everyday Preferences of Nursing Home Residents: Perceived Choice and Satisfaction with Care. The Gerontologist. (Advance online publication): 1-8.

Berlowitz, D. R., Du, W., Kazis, L., & Lewis, S. (1995). Health-related quality of life of nursing home residents: Difference in patient and provider perceptions. Journal of the American Geriatric Society, 43, 799-802.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Gesell, S.B. (2001). A measure of satisfaction for the assisted-living industry. Journal for Healthcare Quality, 23(2), 16-25.

Hall J, Milburn M, Roter D, Daltroy L. Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 1998;17(1):70–75
Hudak, P. L. & J.G. Wright. (2000). The Characteristics of Patient Satisfaction Measures. Spine 25 (24): 3167-3177.

Institute of Medicine (2001). Improving the Quality of Long-Term Care, National Academy Press, Washington, D.C., 2001.

Kajonius, P. & Kazemi, A. (2016). Advancing the Big Five of user-oriented care and accounting for its variations. International Journal of Health Care Quality Assurance. 29(2): 162 – 176.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

Omnibus Budget Reconciliation Act (OBRA) of 1987. (1987, December 22). Public Law 100-203. Subtitle C: Nursing Home Reform.

Riccio, P.A. (2000). Quality Evaluaiton of home nursing care: Perceptions of patients, physicians, and nurses. Nursing Administration Quarterly 24(3): 43-52.

Shippee, T.P., Henning-Smith, C., Kane, R.L, & Lewis, T. (2015). Resident- and Facility-Level Predictors of Quality of Life in Long-Term Care. The Gerontologist. 55(4):643-655.

Uman, C & Urman, H. (1997). Measuring consumer satisfaction in nursing home residents. Nutrition 13: 705-707.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

The consumer movement has fostered the notion that patient evaluations should be an integral component of health care. Patient satisfaction, which is one form of patient evaluation, became an essential outcome of health care widely advocated for use by researchers and policy makers. Managed care organizations, accreditation and certification agencies, and advocates of quality improvement initiatives, among others, now promote the use of satisfaction surveys. For example, satisfaction information is included in the Health Plan Employer Data Information Set (HEDIS), which is used as a report card for managed care organizations (NCQA, 2016).

Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes. A 2014 systematic review and meta-analysis of randomized controlled trials, in which the patient-provider relationship was systematically manipulated and tracked with health care outcomes, found a small but statistically significant positive effect of the patient-provider relationship on health care outcomes (Kelly et al., 2014). This finding aligns with other studies that show a link between patient satisfaction and the following health-related behaviors:

1.Keeping follow-up appointments (Hall, Milburn, Roter, & Daltroy, 1998);
2.Disenrollment from health plans (Allen & Rogers, 1997); and,
3.Litigation against providers (Penchansky & Macnee, 1994).

The positive effect of person-centered care and patient satisfaction is not precluded from skilled nursing facilities. A 2013 systematic review of studies on the effect of person-centered initiatives in nursing facilities, such as the Eden Alternative, found person-centered care associated with psychosocial benefits to residents and staff, notwithstanding variations and limitations in study designs (Brownie & Nancarrow, 2013).

From the nursing facility and provider perspective, there are numerous ways to improve patient satisfaction. One study found conversations regarding end-of-life care options with family members improve overall satisfaction with care and increase use of advance directives (Reinhardt et al., 2014). Another found an association between improving symptom management of nursing home residents with dementia and higher satisfaction with care (Van Uden et al., 2013). Improvements in a nursing home food delivery system also were associated with higher overall satisfaction and improved resident health (Crogan et al., 2013). The advantage of the CoreQ: Short Stay Discharge questionnaire is it is broad enough to capture patient dissatisfaction on various provided services and signal to providers to drill down and discover ways of improving the patient experience at their facility.

Specific to the CoreQ: Short Stay Discharge questionnaire, the importance of the satisfaction areas assessed were examined with focus groups of residents and family members. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. Table 1c.5 in the appendix shows the score of the importance for question included in the CoreQ: Short Stay Discharge questionnaire. The overall ranking used was 10=Most important and 1=Least important. The final four questions included in the measure had average scores ranging from 9.35 to 9.69; this clearly shows that the respondents value the items used in the CoreQ: Short Stay Discharge measure.

Allen HM, & Rogers WH. (1997). The Consumer Health Plan Value Survey: Round Two. Health Affairs. 1997;16(4):156–66.

Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. Clinical Interventions In Aging. 8:1-10.

Crogan, N.L., Dupler, A.E., Short, R., & Heaton, G. (2013). Food choice can improve nursing home resident meal service satisfaction and nutritional status. Journal of Gerontological Nursing. 39(5):38-45.

Hall J, Milburn M, Roter D, Daltroy L (1998). Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 17(1):70–75.

Kelley J.M., Kraft-Todd G, Schapira L, Kossowsky J, & Riess H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and metaanalysis of randomized controlled trials. PLoS One. 9(4): e94207.

Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care was generally high during 2005-09, with some variability across facilities. Health Affairs. 32(8):1416-25.

Lin, J., Hsiao, C.T., Glen, R., Pai, J.Y., & Zeng, S.H. (2014). Perceived service quality, perceived value, overall satisfaction and happiness of outlook for long-term care institution residents. Health Expectations. 17(3):311-20.

National Committee for Quality Assurance (NCQA) (2016). HEDIS Measures. http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures.aspx. Accessed March 2016.

Penchansky and Macnee, (1994). Initiation of medical malpractice suits: a conceptualization and test. Medical Care. 32(8): pp. 813–831.

Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. Journal Of Social Work In End-Of-Life & Palliative Care. 10(2):112-26.

Van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. International Psychogeriatrics. 25(10):1697-707.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Not Applicable

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not Applicable

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The measure assesses the number of patients who are discharged from a SNF, within 100 days of admission, who are satisfied. The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the four questions on the CoreQ: Short Stay Discharge questionnaire.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The CoreQ: Short Stay Discharge questionnaire should be administered to discharge patients within 2 weeks of their discharge date. Patients must respond to the questionnaire within 2 months of receiving the questionnaire.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator includes all of the patients who were discharged within 100 days of admission and had an average response =>3 on the CoreQ: Short Stay Discharge questionnaire.

The calculation of the individual patient's average satisfaction score is done in the following manner:
-A numeric score is associated with each response scale option on the CoreQ: Short Stay Discharge questionnaire (that is, Poor=1, Average=2, Good=3, Very Good=4, and Excellent=5).
-The following formula is utilized to calculate the individual's average satisfaction score: [Numeric Score Question 1 + Numeric Score Question 2 + Numeric Score Question 3 + Numeric Score Question 4]/4
-The number of respondents whose average satisfaction score >=3 are summed together and function as the numerator.

For patients with one missing data point (from the four items included in the questionnaire) imputation is used (representing the average value from the other three available responses). Patients with more than one missing data point, are excluded from the analyses (i.e., no imputation will be used for these patients). Imputation details are described further below (S.22).

No risk-adjustment is used (See S.18).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator includes all of the patients that are admitted to the SNF, regardless of payor source, for post-acute care, that are discharged within 100 days; who receive the survey (e.g. people meeting exclusions do not receive a questionnaire) and who respond to the CoreQ: Short Stay Discharge questionnaire within the time window (See: S.5).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population includes all of the individuals who respond to the CoreQ: Short Stay Discharge questionnaire within the time window (See: S.5).

The data is collected over a maximum 6 month time window. A shorter period can be used if the sample size (125) meets the specifications described below. The questionnaire is administered to discharged patients within 2 weeks of their discharge date. The discharge date is identified from nursing facility records (e.g., MDS, wherein a discharge MDS record is created that includes a discharge date). Note, the questionnaire must be administered after the patient is discharged and not on the day of the discharge. Patients must respond to the CoreQ: Short Stay Discharge questionnaire within 2 months of receiving the questionnaire.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclusions used are made at the time of sample selection and include:

(1) Patients who died during their SNF stay;

(2) Patients discharged to a hospital, another SNF, psychiatric facility, inpatient rehabilitation facility or long term care hospital;

(3) Patients with court appointed legal guardian for all decisions;

(4) Patients discharged on hospice;

(5) Patients who left the nursing facility against medical advice (AMA);

(6) Patients who have dementia impairing their ability to answer the questionnaire defined as having a BIMS score on the MDS as 7 or lower. [Note: we understand that some SNCCs may not have information on cognitive function available to help with sample selection. In that case, we suggest administering the survey to all residents and assume that those with cognitive impairment will not complete the survey or have someone else complete on their behalf which in either case will exclude them from the analysis.]

(7) Patients who responded after the two month response period; and

(8) Patients whose responses were filled out by someone else.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Individuals are excluded based on information from the admission Minimum Data Set (MDS) 3.0 assessment.
(1) Patients who die: This is recorded in the MDS as Deceased (A2100 = 08).

(2) Patients who were discharged to a hospital, another SNCC, psychiatric facility, Inpatient Rehabilitation Facilities (IRF), or MR/DD facility: This is recorded in the MDS as Discharge to hospital (A2100 = 03); another SNCC (A2100 = 02); psychiatric facility (A2100 = 04); Inpatient Rehabilitation Facilities (A2100 = 05); ID/DD facility (A2100 = 06).

(3) Patients with Court appointed legal guardian for all decisions as identified from the nursing facility health

information system.

(4) Patients on hospice: This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").

(5) Patients who left the nursing facility against medical advice (AMA) as identified from nursing facility health information systems.

(6) Patients with a BIMS score on the MDS as 7 or lower. This is recorded in the MDS as C0500 <= 7.

(7) Patients who respond after the two month response period.

(8) Patients whose responses were filled out by somebody other than him/herself, as identified by the additional questions on the questionnaire.

Surveys returned as undeliverable are also excluded from the denominator.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
No stratification is used (see below).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not Applicable

S.16. **Type of score:**
Other (specify):
If other: Non-weighted score. Score is a percentage.

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event,*

or outcome; aggregating data; risk adjustment; etc.)

1.Identify SNF patients that are discharged within 100 days after admission

a.Calculate the duration of the SNF stay [MDS discharge date (A2000) - MDS admission date (A1900)] to determine if it is = 100 days.

2.Take the patients that have a SNF stay of = 100 days and exclude the following:

a.Patients who died; patients discharged to a hospital; patients with Court appointed legal guardian for all decisions; patients with hospice; patients who left the nursing facility against medical advice (AMA), and patients with a BIMS score of less than 7 do not receive that survey as a result of the exclusions (described in detail above).

i.Patients who die: This is recorded in the MDS as Die during stay (A2100 = 08)

ii.Patients who were discharged to a hospital, another SNCC, psychiatric facility, Inpatient Rehabilitation Facility, or MR/DD facility (A2100 = 06): This is recorded in the MDS as Discharge to hospital (A2100 = 03); another SNCC (A2100 = 02); psychiatric facility (A2100 = 04); Inpatient Rehabilitation Facility (A2100 = 05); MR/DD facility (A2100 = 06).

iii.Patients with Court appointed legal guardian for all decisions will be identified from nursing facility health information system.

iv.Patients on hospice: This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").

v.Patients who left the nursing facility against medical advice (AMA) will be identified from nursing facility health information systems.

vi.Patients with a BIMS score of 7 or less. This is recorded in the MDS as C0500 <= 7.

3.Administer the CoreQ: Short Stay Discharge questionnaire (See S.25) to these individuals. The questionnaire should be administered to patients discharged within 2 weeks of discharge. Provide individuals 2 months to respond to the survey.

a.Create a tracking sheet with the following columns:

i.Data Administered

ii.Data Response Received

iii.Time to Receive Response ([Date Response Received – Date Administered])

b.Exclude any surveys where Time to Receive Response >2 Months

4.Collect data over a maximum 6 month time window or until 125 consecutive usable surveys are received (See S.21).

5.Exclude responses not completed by the intended recipient (e.g. questions were answered by a friend or family members. It is important to note that cases in which the residents had help with reading the questions, or writing down their responses, are included in the measure, because in these cases the residents answer the questions themselves).

6.Exclude surveys that are returned after two months

7.Combine the CoreQ: Short Stay Discharge questionnaire items to calculate a patient level score. Responses for each item should be given the following scores:

a.Poor = 1,

b.Average = 2,

c.Good = 3,

d.Very good =4 and

e.Excellent = 5.

8.Impute missing data if only one of the four questions are missing data by taking the average of the other questions responses.

9.Exclude any survey with 2 or more survey questions that have missing data.

10.Calculated patient score from usable surveys.
Patient score= (Score for Item 1 + Score for Item 2 + Score for Item 3 + Score for Item 4) / 4.
a.For example, a patient rates their satisfaction on the CoreQ questions as excellent = 5, very good = 4, very good = 4, and good = 3. The resident's total score will be 5 + 4 + 4 + 3 for a total of 16. The patient's total score (16) will then be divided by the number of questions (4), which equals 4. Thus the patients average satisfaction rating is 4.0. This individual would be counted in the numerator since their average score is >3.0.

11.Flag those patients with an average score equal to or greater than 3.0

12.Calculate the CoreQ: Short Stay Discharge measure which represents the percent of patients with average scores of 3.0 or above.
CoreQ: Short Stay Measure= ([number of valid responses with an average score of =3.0] / [total number of valid responses])*100

13.No risk-adjustment is used.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
No sampling is used. No proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
1.Administer that CoreQ: Short Stay Discharge questionnaire to SNF patients discharged within 100 days of admission and who do not fall into one of the exclusions noted below.
a.Identify that SNF patient is discharged within 100 days of admission
i.Calculate the duration of the SNF stay [MDS discharge date (A2000) - MDS admission date (A1900)] to determine if it is = 100 days.
b.Remove individuals with the following exclusions from the sample:
i.Patients who die: This is recorded in the MDS as Die during stay (A2100 = 08)
ii.Patients who were discharged to a hospital, another SNCC, psychiatric facility, Inpatient Rehabilitation Facility, or MR/DD facility (A2100 = 06). This is recorded in the MDS as Discharge to hospital (A2100 = 03); another SNCC (A2100 = 02); psychiatric facility (A2100 = 04); Inpatient Rehabilitation Facility (A2100 = 05); MR/DD facility (A2100 = 06).
iii.Patients with Court appointed legal guardian for all decisions will be identified from nursing facility health information system.
iv.Patients on hospice: This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").
v.Patients who left the nursing facility against medical advice (AMA) will be identified from nursing facility health information system.
vi.Patients with a BIMS score of 7 or lower. This is recorded in the MDS as C0500 <= 7.

2.Administer the CoreQ: Short Stay Discharge questionnaire to patients discharged, within two weeks of discharge (ideally, within one week). The questionnaire should be administered after discharge, not the day of discharge. Optional but not required, reminders or duplicate questionnaires can be administered to patients to help increase response rate.

3.Instruct individuals that they must respond to the survey within two months.

4.Collect the responses continuously for all eligible discharges. The maximum time period for data collection is 6 months. However, a SNF may optionally stop data collection if they consecutively receive =125 usable surveys and calculate the measure.

5.A minimum response rate of 30% needs to be achieved for results to be reported for a SNF.
a.The response rate is calculated as the number of valid returned questionnaires divided by the number of questionnaires administered. Those returned as undeliverable are excluded as well as those completed by another person on behalf of the patient and those with missing data on 2 or more of the 4 questions.

6.Regardless of response rate, SNFs must also achieve a minimum number of 20 usable questionnaires (e.g. denominator). If after 6 month, less than 20 usable questionnaires are received than a facility level satisfaction measure cannot be reported.

7.All the questionnaires that are received (other than those with more than one missing value; or those returned as undeliverable; or those returned after two months; or those completed by another person) must be used in the calculations.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data was uncommon in the CoreQ: Short Stay Discharges questionnaire testing (<13% of any one of the 4 items). For patients with one missing data point (from the 4 items included in the CoreQ: Short Stay Discharge questionnaire) imputation is utilized (representing the average value from the other available data points). Patients with more than one missing data point, are excluded from the analyses (i.e., no imputation will be used).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Healthcare Provider Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The collection instrument is the CoreQ: Short Stay Discharge questionnaire and Resident Assessment Instrument Minimum Data Set (MDS) version 3.0.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
CoreQ_Short_Stay_Testing_Final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: CoreQ: Short Stay Discharge Measure
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***		☒ Outcome (*including PRO-PM*)
☐ Cost/resource		☐ Process
☐ Efficiency		☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance

score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are

different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: CoreQ: Short Stay Discharge questionnaire	☒ other: CoreQ: Short Stay Discharge questionnaire, Pilot CoreQ: Short Stay Discharge questionnaire, Nursing Home Compare, and CASPER

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities*

being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).
Data utilized for testing came from CoreQ: Short Stay Discharge questionnaire. To validate the measure; we also utilized CASPER Quality Indicators and data form Nursing Home Compare. Additionally, Pilot CoreQ: Short Stay Discharge questionnaire containing an extended list of questions included on the CoreQ: Short Stay Discharge questionnaire was utilized for reliability and validity testing.

1.3. What are the dates of the data used in testing? Click here to enter date range
June, 2014-September, 2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: Individual Resident

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
The testing and analysis included four data sources (Table 1.5 below):

1. Reliability and validity testing of the Pilot CoreQ: Short Stay Discharge questionnaire was examined using responses from 853 patients from a national sample of facilities.
2. Validity testing of the Pilot CoreQ: Short Stay Discharge questionnaire was examined using responses from 100 patients from the Pittsburgh area.
3. CoreQ: Short Stay Discharge measure was examined using 282 facilities and included responses from 10,319 patients. These facilities were located across multiple states.
4. In addition, patient-level sociodemographic (SDS) variables were examined using a sample of 1012 patients in nursing facilities in Massachusetts. This included 121 facilities.

Table 1.5: Measured Entities

Data Source	Average Number of	Average Daily	Average Monthly Number of New	Sample Size of Patients

	Licensed Beds	Census	Patients	(N)
Listed #1 (above)	122	112	37	853
Listed #2 (above)	202	188	49	100
Listed #3 (above)	135	108	34	10,319
Listed #4 (above)	140	133	29	1,012

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Patient Level of Analysis

Data was used from the CoreQ: Short Stay Discharge questionnaire. The questionnaire was mailed to all patients discharged within 2 weeks of their discharge date (with the exclusions described in the Specification section). The testing and analysis included:

1. The Pilot CoreQ: Short Stay Discharge questionnaire was examined using responses from 853 patients from a national sample of facilities.
2. Validity testing of the Pilot CoreQ: Short Stay Discharge questionnaire was examined using responses from 100 patients from the Pittsburgh area.
3. CoreQ: Short Stay Discharge measure was examined using 282 facilities and included responses from 10,319 patients. These facilities were located across multiple states.
4. In addition, patient-level sociodemographic (SDS) variables were examined using a sample of 1012 patients in nursing facilities in Massachusetts. This included 121 facilities.

The descriptive characteristics of the residents are given in the following table that includes information from all of the data used (the education level and race information comes only from the sample described above with 1012 respondents, as this data was not collected for the other samples).

Table 1.6: Descriptive Characteristics of Patients Included in the Analysis (all samples pooled)

DEMOGRAPHICS		Percent
How long were you a resident at this facility?	<1 Month	60.88%
	1-3Months	34.59%

	3-6Months	2.89%
Are you male or female?	Male	39%
	Female	61%
What year were you born?	Average	1936
What is the highest grade or level of school that you have completed?	Some HS	15%
	HS or GED	41%
	Some College/ 2yr Degree	23%
	4yr College Degree	11%
	>4yr College Degree	10%
Are you of Hispanic or Latino origin or descent?	Yes	2%
	No	98%
What is your race?	White	86%
	Black	13%
	Asian	1%
	Native Hawaiian	0%
	American Indian	0%

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

We conducted two levels of testing in the development of the CoreQ: Short Stay Discharge measure. The first focused on testing (e.g., reliability, validity, exclusions) of the CoreQ: Short Stay Discharge questionnaire. The first source of data (pilot data) was utilized in developing and choosing the items to be included in the CoreQ: Short Stay Discharge questionnaire. This included using a questionnaire with 22 items. Below we call this the Pilot CoreQ: Short Stay Discharge questionnaire.

Once the CoreQ: Short Stay Discharge questionnaire was developed, a second source of data was used to test the validity of the CoreQ: Short Stay Discharge measure (i.e., facility and summary score validity).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient

(e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The following patient-level sociodemographic variables were available for analysis. For the distribution of these categories, see Table 1.6 above.

- Age
  - Exact date of birth
- Sex
  - Male
  - Female
- Highest level of education
  - Some high school, but did not graduate
  - High school graduate or GED
  - Some college or 2 year degree
  - 4 year college graduate
  - More than 4 year college degree
- Hispanic Descent
  - Yes
  - No
- Race
  - White
  - Black or African American
  - Asian
  - Native Hawaiian or other Pacific Islander
  - American Indian or Alaskan Native.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

We measured reliability at the: (1) data element level; (2) the person/questionnaire level; and, (3) at the measure (i.e., facility) level. More detail of each analysis follows.

(1) DATA ELEMENT LEVEL

To determine if the CoreQ: Short Stay Discharge questionnaire data elements were repeatable (i.e. producing the same results a high proportion of the time when assessed in the same population in the same time period) we re-administered the questionnaire to patients 1 month after the submission of their first survey. The Pilot CoreQ: Short Stay Discharge questionnaire had responses from 853 patients; we re-administered the survey to 100 patients. The re-

administered sample was a sample of convenience as they represented patients from the Pittsburgh area (the location of the team testing the questionnaire). To measure the agreement, we calculated first the distribution of responses by question in the original round of surveys, and then again in the follow-up surveys (they should be distributed similarly); and second, calculated the correlations between the original and follow-up responses by question (they should be highly correlated).

(2) PERSON/QUESTIONNAIRE LEVEL

Having tested whether the *data elements* matched between the pilot responses and the re-administered responses, we then examined whether the *person-level* results matched between the Pilot CoreQ: Short Stay Discharge questionnaire responses and their corresponding re-administered responses. In particular, we calculated the percent of time that there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re- administered response was poor, average, good, very good or excellent.

(3) MEASURE (FACILITY) LEVEL

Last, we measured stability of the facility-level measure when the facility's score is calculated using multiple "draws" from the same population. This measures how stable the facility's score would be if the underlying patients are from the same population but are subject to the kind of natural sample variation that occurs over time. We did this by bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score calculated on the Pilot CoreQ: Short Stay Discharge questionnaire sample.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

(1) DATA ELEMENT LEVEL

Table 2a2.3.a shows the four CoreQ: Short Stay Discharge questionnaire items, and the response per item for both the pilot survey of 853 patients and the re-administered survey of 100 patients. The responses in the pilot survey are not statistically significant from the re-administered survey. This shows that the data elements were highly repeatable and produced the same results a high proportion of the time when assessing the same population in the same time period.

Table 2a2.3.a: CoreQ: Short Stay Discharge Questionnaire Responses from the Pilot and Re-administered Surveys

Questionnaire Item	Response	Percent [Pilot Survey (N=853)]	Percent [Re-Mailed Survey (N=100)]
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	10%	11%
	Average	10%	9%
	Good	15%	13%
	Very Good	33%	35%

	Excellent	33%	33%
Questionnaire Item	**Response**	**Percent [Pilot Survey (N=853)]**	**Percent [Re-Mailed Survey (N=100)]**
2. Overall, how would you rate the staff?	Poor	4%	4%
	Average	10%	10%
	Good	17%	16%
	Very Good	40%	42%
	Excellent	30%	29%
3. How would you rate the care you received?	Poor	5%	5%
	Average	12%	13%
	Good	18%	18%
	Very Good	37%	36%
	Excellent	28%	27%
4. How would you rate the discharge process?	Poor	8%	8%
	Average	12%	13%
	Good	20%	20%
	Very Good	34%	33%
	Excellent	26%	25%

Table 2a2.3.b shows the average of the percent agreement from the first survey score to the second survey score for each item in the CoreQ: Short Stay Discharge questionnaire. This shows very high levels of agreement.

Table 2a2.3.b: Average Percent Agreement between 1st and 2nd Administered Surveys

Questionnaire Item	Percent Agreement
5. In recommending this facility to your friends and family, how would you rate it overall?	96.8%
6. Overall, how would you rate the staff?	97.8%
7. How would you rate the care you receive?	98.2%
8. How would you rate the discharge process?	98.2%

(2) PERSON/QUESTIONNAIRE LEVEL

Table 2a2.3.c shows the CoreQ: Short Stay Discharge questionnaire items, and the agreement in response per item for both the pilot survey of 853 patients compared with the re- administered survey of 100 patients. The person-level responses in the pilot survey are not statistically significant from the re- administered survey. This shows that a high percent of time there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re- administered response was poor, average, good, very good or excellent. Table 2a2.3.d shows the agreement between the pilot and re- administered responses. In summary, 98% or more of the re- administered responses agreed with their corresponding pilot responses, in terms of whether or not they were rated in the categories of poor or average or good, very good or excellent.

Table 2a2.3.c: Average Percent Agreement between Responses per Item for the Pilot Survey and Re- administered Survey

Questionnaire Item	Response	Percent Person-Level Agreement in Response for the Pilot Survey (N=853) vs. Re-administered Survey (N=100)
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	96%
	Average	96%
	Good	95%
	Very Good	98%
	Excellent	99%
2. Overall, how would you rate the staff?	Poor	99%
	Average	98%
	Good	98%
	Very Good	96%
	Excellent	98%
3. How would you rate the care you received?	Poor	99%
	Average	99%
	Good	98%
	Very Good	97%
	Excellent	98%
4. How would you rate the discharge process?	Poor	99%
	Average	97%
	Good	98%

	Very Good	99%
	Excellent	98%

Table 2a2.3.d: Average Percent Agreement between Response Options for the Pilot Survey and Re- administered Survey

		Re- administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
Pilot Response	Poor (1) or Average (2)	98.5%	98%
	Good (3), Very Good (4), or Excellent (5)	98.5%	99%

(3) MEASURE (FACILITY) LEVEL

After having performed the 10,000-repetition bootstrap, 17.82% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample, 38.14% were within 3 percentage points, 61.05% were within 5 percentage points, and 87.05% were within 10 percentage points.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

In summary, the measure displays a high degree of element-level, questionnaire-level, and measure (facility)-level reliability. First, the CoreQ: Short Stay Discharge questionnaire data elements were highly repeatable, with pilot and re-administered responses agreeing between 94% and 97% of the time, depending on the question. That is, this produced the same results a high proportion of the time when assessed in the same population in the same time period. Second, the questionnaire level scores were also highly repeatable, with pilot and re-administered responses agreeing 98% of the time. Third, a facility drawing patients from the same underlying population only varied modestly. The 10,000-repetition bootstrap results showed that the CoreQ: Short Stay Discharge measure scores from the same facility are very stable, given the minimum sample size of 20 we set for this measure; and the maximum sample size of 196.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

In the development of the CoreQ: Short Stay Discharge questionnaire, four sources of data were used to perform three levels of validity testing. Each is described further below. The first source of data (convenience sampling) was used in developing and choosing the format to be utilized in the CoreQ: Short Stay Discharge questionnaire (i.e., response scale). The second source of data was pilot data collected from 865 patients (described below). This data was used in choosing the items to be used in the CoreQ: Short Stay Discharge questionnaire. The third source of data (collected from 285 facilities described in Section 1.5) was used to examine the validity of the CoreQ: Short Stay Discharge measure (i.e., facility and summary score validity).

Thus, the following sections describe this validity testing:
1. Validity testing of the questionnaire format used in the CoreQ: Short Stay Discharge questionnaire;
2. Testing the items for the CoreQ: Short Stay Discharge questionnaire;
3. To determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Short Stay Discharge measure);
4. Validity Testing for the CoreQ: Short Stay discharge measure.

In summary, the overall intent of these analyses was to determine if a subset of items could reliably be used to produce an overall indicator of satisfaction.

1. *Validity Testing for the Questionnaire Format used in the CoreQ: Short Stay Discharge Questionnaire*

A. The face validity of the domains used in the CoreQ: Short Stay Discharge questionnaire was evaluated via a literature review. The literature review was conducted to examine important areas of satisfaction for long term care residents. The research team examined 12 commonly used satisfaction surveys and reports to determine the most valued satisfaction domains. These surveys were identified by completing internet searches in PubMed and Google. Key terms that were searched included "resident satisfaction, long-term care satisfaction, and elderly satisfaction".

B. The face validity of the domains was also examined using patients. The overall ranking used was 1=Most important and 22=Least important. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region.

C. The face validity of the Pilot CoreQ: Short Stay Discharge questionnaire response scale was also examined. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. The percent of respondents that stated they "fully understood" how the response scale worked, could complete the scale, AND in cognitive testing understood the scale was used.

D. The Flesch-Kinkaid scale (Streiner & Norman, 1995) was used to determine if respondent correctly understood the questions being asked (Streiner, D. L. & Norman, G.R., 1995).

2. Testing the Items for the CoreQ: Short Stay Discharge Questionnaire

The analyses above were performed to provide validity information on the format in the CoreQ: Short Stay Discharge questionnaire (i.e, domains and format). The second series of validity testing was used to further identify items that should be included in the CoreQ: Short Stay Discharge questionnaire. This analysis was important, as all items in a satisfaction measure should have adequate psychometric properties (such as low basement or ceiling effects). For this testing, a Pilot version of the CoreQ: Short Stay Discharge questionnaire survey was administered consisting of 22 items (N= 853 patients). The testing consisted of:

A. The Pilot CoreQ: Short Stay Discharge questionnaire items performance with respect to the distribution of the response scale and with respect to missing responses.

B. The intent of the pilot instrument was to have items that represented the most important areas of satisfaction (as identified above) and to be parsimonious. Additional analyses were used to eliminate items in the Pilot instrument. More specifically, analyses such as exploratory factor analysis (EFA) were used to further refine the pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items.

3. Determine if a Sub-Set of Items Could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Short Stay Discharge measure).

The CoreQ: Short Stay Discharge is meant to represent overall satisfaction with as few items as possible. The testing given below describes how this was achieved.

A. To support the construct validity (i.e. that the CoreQ items measured a single concept of "satisfaction") we performed a correlation analysis using all items in the instrument.

B. In addition, using all items in the instruments a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item was explored.

4. Validity Testing for the Core Q: Short Stay Discharge Measure.

The overall intent of the analyses described above was to identify if a sub-set of items could reliably be used to produce an overall indicator of satisfaction, the CoreQ: Short Stay Discharge questionnaire. Further testing was conducted to determine if the 4 items in the CoreQ: Short Stay Discharge questionnaire were a reliable indicator of satisfaction.

A. To determine if the 4 items in the CoreQ: Short Stay Discharge questionnaire were a reliable indicator of satisfaction, the correlation between these four items in the CoreQ: Short Stay Discharge Measure and all of the items on the Pilot CoreQ instrument was conducted.

B. We performed additional validity testing of the facility-level CoreQ: Short Stay Discharge measure by measuring the correlations between the CoreQ: Short Stay Discharge measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, ii) several other quality metrics from Nursing Home Compare, iii) risk adjusted discharge to community measure and iv) risk adjusted PointRight® Pro 30™ Rehospitalizations. If the CoreQ Short Stay Discharge scores correlate negatively with the measures that decrease as they get better, and positively with the measures that increase as they get better, then this supports the validity of the CoreQ Short Stay Discharge measure.

Streiner, D. L. & Norman, G.R. 1995. Health measurement scales: A practical guide to their development and use. 2nd ed. New York: Oxford.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Short Stay Discharge questionnaire

A. The face validity of the Domains used in the CoreQ: Short Stay Discharge questionnaire was evaluated via a literature review (described in 2b2.2). Specifically, the research team examined the surveys and reports to identify the different domains that were included. The research team scored the domains by simply counting if an instrument included the domain. Table 2b2.3.a gives the domains that were found throughout the search, as their respective score. An example is the domain food, this was used in 11 out of the 12 surveys. (Note: food was not ultimately included in the final CoreQ Short Stay Discharge because correlation and factor analysis showed that it added little to the survey when the overall question, i.e. CoreQ Question 1 was used). An interpretation of this finding would be that items addressing food are extremely important in satisfaction surveys. These domains were used in developing the pilot CoreQ: Short Stay Discharge questionnaire items.

Table 2b2.3.a: Survey Domain Score out of 12

Domain	Score out of 12	Domain	Score out of 12
Food	11	Spiritual	4
Activities	10	Confidence in Caregivers	3
Administration	10	Language and Communication	3
Clinical Care	10	Personal Suite	3
Staff Interaction	10	Therapy	3
Choice and Decision Making	9	Care Access	2
Facility Environment	9	Case Manager	2
Security and Safety	9	Comfort	2
Overall	8	Maintenance	2

Staff Overall	7	Move In	2
Autonomy and Privacy	6	Non-Clinical Staff Services	2
Housekeeping	6	Transitions	2
Personal Care	6	Transportation	2
Recommend facility	6	Emergency Response	1
Resident to Resident Friendships	5	Finances	1
Family Involvement	4	Time	1
Resident to Staff Friendships	4	Trust	1

B. The face validity of the domains was also examined using patients (described above). The following abbreviated table shows the rank of importance for each group of domains. The overall ranking used was 1=Most important and 22=Least important. The ranking of the 4 areas used in the CoreQ: Short Stay Discharge questionnaire are shown in Table 2b2.3.b.

Table 2b2.3.b: Average Ranking of CoreQ: Average Ranking of CoreQ: Short Stay Discharge Questionnaire Items

Domain / Question	Average Rank
OVERALL (In recommending this facility to your friends and family, how would you rate it overall?)	2
STAFF (Overall, how would you rate the staff?)	1
CARE (How would you rate the care you received?)	3
DISCHARGE (How would you rate how well your discharge needs were met?)	5

C. The face validity of the pilot CoreQ: Short Stay Discharge questionnaire response scale was also examined (described above). Table 2b2.3.c gives the percent of respondents that stated they fully understood how the response scale worked, could complete the scale, AND in cognitive testing understood the scale.

Table 2b2.3.c: Resident Understanding of Response Scale

Scale Format	Residents
Yes – No	100%
Yes – Somewhat – No	100%
Always – Usually – Sometimes –Never	100%
Very happy – Somewhat happy – Unhappy	100%
Excellent – Good – Fair – Poor	100%

Very Good – Good – Average – Poor – Very Poor	100%
Very Satisfied – Satisfied – Neither Satisfied or Dissatisfied – Dissatisfied – Very Dissatisfied	100%
4 Point Satisfaction Scale (1=Very unsatisfied, 2=Unsatisfied, 3=Neutral, 4=Satisfied)	100%
5 Point Likert Scale (1=Poor, 2=Average, 3=Good, 4=Very Good, 5=Excellent)	100%
Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree	95%
5 Point Importance Scale (1=Very important, 5=Very unimportant)	95%
5 Point Expectancy Scale (1=Not met, 2=Nearly met, 3=Met, 4=Exceeded, 5=Far exceeded expectations)	90%
10 Point Satisfaction Scale (1=Poor, 10=Excellent)	90%
8 Point Satisfaction Scale (1=Very dissatisfied, 2=Dissatisfied, 3=Somewhat dissatisfied, 4=Neither satisfied nor dissatisfied, 5=Somewhat satisfied, 6=Satisfied, 7=Very satisfied, 8=No response)	85%

Note: Highlighted cell represents the scale used in the CoreQ.

D. The CoreQ: Short Stay Discharge questionnaire was purposefully written using simple language. No *a priori* goal for reading level was set, however a Flesch-Kinkaid scale score of six, or lower, is achieved for all questions.

2. Testing the Items for the CoreQ: Short Stay Discharge Questionnaire

A. The pilot CoreQ: Short Stay Discharge questionnaire items are shown below. Table 2b2.3.d in the appendix shows that the items performed well with respect to the distribution of the response scale and with respect to missing responses.

B. Using all items in the instruments (excluding the global item Q1 ("How would you rate the facility?")) exploratory factor analysis (EFA) was used to evaluate the construct validity of the measure. The Eigenvalues from the principal factors (unrotated) are presented. Sensitivity analyses using principal factors and rotating provide highly similar findings.

Table 2b2.3.e: Eigenvalues for Principle Factors

	Short-Stay Resident
Factor 1	9.61
Factor 2	0.37

3. *Determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Short Stay Discharge Measure).*

A. To support the construct validity that the idea that the CoreQ items measured a single concept of "satisfaction" – we performed a correlation analysis using all items in the instrument. The analysis identifies the pairs of CoreQ items with the highest correlations. The highest correlations are shown in Table 2b2.3.f. Items with the highest correlation are potentially providing similar satisfaction information. Note, the table provides 7 sets of correlations, the analysis was conducted examining all possible correlations between items. Because items with the highest correlation were potentially providing similar satisfaction information they could be eliminated from the instrument.

Table 2b2.3.f: CoreQ: Short Stay Discharge Questionnaire Item Correlations

	Short-stay
Highest Correlation	Q8-Q6 (.841)
Next highest Correlation	Q10-Q9 (.842)
Next highest Correlation	Q17-Q20 (.822)
Next highest Correlation	Q6-Q2 (.814)
Next highest Correlation	Q15-Q6 (.804)
Next highest Correlation	Q13-Q10 (.814)
Next highest Correlation	Q9-Q2 (.818)

RESULT = ITEMS TO DROP

B. In addition, using all items in the instrument a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item is shown in table 2b2.3.g. Chronbach's alpha measures the internal consistency of the values entered into the factor analysis, where a value of 0.7 or higher is generally considered acceptably high. The additional item(s) is considered best in the sense that it is most highly correlated with the existing item, and therefore provides little additional information about the same construct. So this analysis was also used to eliminate items. Note, the table again provides 7 sets of correlations, the analysis was conducted examining all possible correlations between items.

Table 2b2.3.g: Secondary Correlation Analysis of CoreQ: Short Stay Discharge Questionnaire Items

	Short-stay
Q1 + last satisfaction item ADD	Q10 (.941) Q6 (.937) Q2 (.931)
Q1 + ADD	Q2 + Q6 (.934) Q10 + Q9 (.930)

ADD	Q9 + Q8 (.921)
Q1 +	Q10 + Q6 (.934)
ADD	Q10 + Q9 (.934)
ADD	Q9 + Q6 (.930)

Thus, using the correlation information and factor analysis 4 items representing the CoreQ: Short Stay Discharge questionnaire were identified.

4. ***Validity testing for the Core Q: Short Stay Discharge Measure***

The overall intent of the analyses described above was to identify if a sub-set of items could reliably be used to produce an overall indicator of satisfaction, the CoreQ: Short Stay Discharge questionnaire.

A. The items were all scored according to the rules identified elsewhere. The same scoring was used in creating the 4 item CoreQ: Short Stay Discharge questionnaire summary score and the satisfaction score using the Pilot CoreQ: Short Stay Discharge questionnaire. The correlation was identified as having a value of 0.94.

That is, the correlation score between the final "CoreQ: Short Stay Discharge Measure" and all of the 22 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 4 items or the 22 item Pilot instrument.

B. We performed additional validity testing of the facility-level CoreQ: Short Stay Discharge measure by measuring the correlations between the CoreQ: Short Stay Discharge measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, ii) several other quality metrics from Nursing Home Compare, iii) risk-adjusted Discharge to Community Measure [NQF# 2858] and iv) risk-adjusted PointRight® Pro 30™ Rehospitalizations [NQF# 2375].This score should be associated with better quality in the SNF. Therefore, we hypothesize that for each facility in the sample there is a positive correlation with other quality indicators.

(i) Relationship with CASPER Quality Indicators

Certification and Survey Provider Enhanced Reporting (CASPER) contains data collected as part of state/federal nursing home inspections. In short, nursing facilities that accept residents with Medicare and/or Medicaid payments are surveyed; this includes most (i.e., 97% [15,000 facilities]) nursing homes in the U.S. The survey process occurs approximately yearly, and includes the recording of many quality characteristics of the nursing home. The most commonly used CASPER quality indicators are restraint use, pressure ulcers, catheter use, antipsychotic use, antidepressant use, antianxiety use, and, use of hypnotics in SNFs.

In addition, when a SNF is determined not to meet a certification minimum standard a deficiency citation is issued. These deficiency citations are also commonly used in the analyses of the quality of SNFs. Approximately 180 deficiency citations exist and are grouped into 16 categories. These 16 categories group similar areas together. They were developed by CMS and have considerable face validity; although, one limitation of using these categories is that they were not defined using empirical estimation (such as factor analysis).

Table 2b2.3.h: CoreQ: Short Stay Discharge Correlation with Quality Metrics

Quality Indicator	Correlation with Satisfaction Summary Score	P-Value
Any Deficiency Citations	-0.11	0.07
Physical Restraint Use	-0.01	0.91
Pressure ulcers	-0.22	<0.01
Catheterized	-0.04	0.56
Antipsychotic medications	-0.06	0.32
Antidepressant medications	0.13	0.03
Antianxiety medications	0.08	0.19
Hypnotic medications	0.04	0.46

(ii) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings and staffing levels

Nursing Home Compare (NHC) is a nursing home report card. After several years of pilot testing, the Centers for Medicare and Medicaid Services (CMS) released this report card on the world-wide web in November of 2002. Briefly, Nursing Home Compare provides information for facility location, structural factors (such as ownership), and staffing characteristics (such as registered nurse [RN] staffing levels). Most significantly, standardized quality information is presented in what are called Quality Measures (QMs). These are calculated from MDS information.

At the time period of for this study (i.e., 2014) CMS reported on 19 measures – these are called the core Quality Measures. The Quality Measures address specific areas of resident care, 5 are for short-stay residents and 14 are for long-stay residents. Long-stay measures are for those residents staying at a facility 3 months or more and short-stay measures are for residents staying at a facility less than 3 months. The short-stay measures are most pertinent to the CoreQ: Short Stay Discharge questionnaire; therefore, these were used in the analyses. These are the percent of residents: with delirium; with moderate to severe pain; and, with pressure sores.

Nursing Home Compare also uses a five-star rating for facilities. This is based on information from the health inspection, direct care staffing, and the MDS quality measures. A five star facility is the highest score and a 1 star facility the lowest score. With respect to staffing, two measures are used: 1) RN hours per resident day; and 2) total staffing hours (RN+ LPN+ nurse aide hours) per resident day.

Table 2b2.3.i: CoreQ: Short Stay Discharge Correlation with Short Stay Quality Measures, Five Star ratings, and staffing levels

Quality Indicator	Correlation with	P-value

	CoreQ: Short Stay Discharge	
Percent of residents with delirium	-0.120	0.30
Percent of residents with moderate to severe pain	-0.138	0.19
Percent of residents with pressure sores	-0.251	0.08
Five-Star rating	0.330	0.07
RN hours per resident day	0.305	0.11

(iii) Relationship with the risk-adjusted Discharge to Community Measure

The risk adjusted Discharge to Community [NQF# 2858] measure determines the percentage of all new admissions from a hospital who are discharged back to the community within 100 days and remain out of any skilled nursing center for the next 30 days. The measure, referring to a rolling year of MDS entries, is calculated each quarter and includes all new admissions to a SNF regardless of payor source. Unsuccessful discharges will result in the resident becoming a long stay resident, which we hypothesize would increase dissatisfaction in SNFs with poor discharge to community rates.

The results of testing for correlation between risk-adjusted discharge to community measure (from 2015q1) and the CoreQ: Short Stay Discharge measure are provided in the table below.

Table 2b2.3.j: Correlation results between the CoreQ: Short Stay Discharge Measure and Risk-adjusted Discharge to Community Measure

CoreQ: Short Stay Discharge	Correlation with Risk-adjusted discharge to community measure	P-Value
Q1: In recommending this facility to your friends and family, how would you rate it overall?	-0.05	0.36
Q2: Overall, how would you rate the staff?	-0.16	0.01
Q3: How would you rate the care you received?	-0.12	0.05
Q4: How would you rate how well your discharge needs were met?	-0.10	0.09
CoreQ: Short Stay Discharge summary score	-0.11	0.06

(iv) Relationship with the risk-adjusted PointRight® Pro 30™ Rehospitalizations

PointRight® Pro 30™ [NQF #2375] is an all-cause, risk adjusted rehospitalization measure. It provides the rate at which all patients (regardless of payer status or diagnosis) who enter skilled nursing facilities (SNFs) from acute hospitals and are subsequently rehospitalized during their SNF stay, within 30 days from their admission to the SNF. Individuals who are rehospitalized after admission are much more likely to become a long stay residents. We hypothesize residents would therefore be more dissatisfied on average in SNFs with high short stay resident rehospitalization rates.

The results of testing for correlation between the risk-adjusted PointRight® Pro 30™ Rehospitalizations measure (from 2015q2) and the CoreQ: Short Stay Discharge measure are provided in the table below.

Table 2b2.3.j: Correlation results between the CoreQ: Short Stay Discharge Measure and Risk-adjusted PointRight® Pro 30™ Rehospitalizations Measure

CoreQ: Short Stay Discharge	Correlation with Risk-adjusted PointRight® Pro 30™ Rehospitalizations measure	P-Value
Q1: In recommending this facility to your friends and family, how would you rate it overall?	-0.23	<0.001
Q2: Overall, how would you rate the staff?	-0.28	<0.001
Q3: How would you rate the care you received?	-0.24	<0.001
Q4: How would you rate how well your discharge needs were met?	0.31	<0.001
CoreQ: Short Stay Discharge summary score	-0.28	<0.001

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Short Stay Discharge Questionnaire

A. The literature review shows that domains used in the Pilot CoreQ: Short Stay Discharge questionnaire items have a high degree of both face validity and content validity.

B. Patients overall rankings, show the general "domain" areas used indicates a high degree of both face validity and content validity.

C. The results show that 100% of residents are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.

D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the items.

2. *Testing the Items for the CoreQ: Short Stay Discharge Questionnaire*

A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

3. Determine if a Sub-Set of Items Could Reliably be Used to Produce an Overall Indicator of Satisfaction (The Core Q: Short Stay Discharge Measure).

A. Using the correlation information of t*he Core Q: Short Stay Discharge questionnaire (22 items)* and the 4 items representing the CoreQ: Short Stay Discharge questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Short Stay Discharge Measure.

A. The correlation of the 4 item CoreQ: Short Stay Discharge measure summary score (identified elsewhere in this document) with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.94.

That is, the correlation score between actual the "CoreQ: Short Stay Discharge Measure" and all of the 22 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 4 items or the 22 item Pilot questions.

This indicates that the CoreQ: Short Stay Discharge instrument summary score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B.

(i) Relationship with CASPER Quality Indicators

The 8 CASPER quality indicators had a low to moderate level of negative correlation with the CoreQ: Short Stay Discharge measure. Those that correlate have a clear conceptual link with short stay, and those that do not are more associated with long stay residents or have unclear conceptual links to short stay customer satisfaction. The CASPER quality indicators that correlate with the CoreQ Short Stay Discharge score are any deficiency citations (-0.11; p=0.07), pressure ulcers (-0.22, p<0.01) and antidepressants (+0.13, p=0.03); those that do not correlate are physical restraints (-0.01, p=0.91), catheterization (-0.04, p=0.56), antipsychotic medications (-0.06, p=0.32), antianxiety medications (0.08, p=0.19), and hypnotic medications (0.04, p=0.46). This testing indicates a moderate degree of construct validity and convergent validity.

(ii) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings and staffing levels

The Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels all had a moderately high levels of correlation and in the direction predicted with the CoreQ: Short-Stay Discharge measure. These correlations range from ± 0.120 to 0.330. The CoreQ: Short-Stay Discharge measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.330 for RN hours per resident-day and 0.305 for total staffing hours per resident day. This testing indicates a high degree of construct validity and convergent validity.

(iii) Relationship with the risk-adjusted Discharge to Community Measure

The risk-adjusted Discharge to community measure was negatively correlated to the CoreQ: Short Stay Discharge measure. The correlations were small ranging from -0.05 to -0.16. This was not as hypothesized which may be related to some SNFs that specialize in long stay, have very low discharge to community rates as admissions do not have a plan to go home.

(iv) Relationship with the risk adjusted PointRight® Pro 30™ Rehospitalizations

The risk-adjusted PointRight® Pro 30™ Rehospitalizations was negatively correlated to the CoreQ: Short Stay Discharge measure. The correlations were modest ranging from -0.22 to -0.31, and all of them were statistically significant at the p-value of 0.05. This is expected because lower rehospitalization rates (an indicator of high quality) are associated with higher satisfaction. This was as hypothesized. This testing indicates a reasonable degree of construct validity and convergent validity.

As noted by Mor and associates (2003, p.41) "there is only a low level of correlation among the various measures of quality." Castle and Ferguson (2010) also show the pattern of findings of quality indicators in nursing facilities is consistently moderate with respect to the correlations identified. Thus, it is not surprising that "very high" levels of correlations were not identified. Nevertheless, some correlation was identified.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
To develop the CoreQ: Short Stay Discharge measure, we convened an expert panel to advise us on aspects such as which exclusions to apply to the measure.

Two sources of data were used to examine the exclusions. The first, included responses from 10,319 patients (Section 1.5). The second exclusion analysis included 100 nursing homes that have used the CoreQ: Short Stay Discharge measure in Massachusetts.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
The expert panel advised us to exclude patients who died, patients who were discharge to a hospital, patients with durable power of attorney for all decisions, patients on hospice, patients with low BIMS scores, and patients who left against medical advice.

These exclusions are often used with satisfaction surveys. Because the exclusions were made we are not able to confirm if the exclusions actually made a difference to the scores, which is why we cannot calculate the mean CoreQ: Short Stay Discharge scores with and without the exclusions. However, we are able to report descriptive statistics regarding the number of exclusions made.

The first, exclusion analysis included responses from 10,319 patients (described elsewhere). The exclusions were tracked and included 1,970 patients (19.1%) discharged to the hospital; 5 (0.05%) discharged to hospice; and, 10 (0.09%) expired. The exclusions of the patients that had left against medical advice or had a durable power of attorney were not tracked in this sample.

The second exclusion analysis included 100 nursing homes and data from the first 1000 patients that were included in this initiative: 791 patients (7.9%) were discharged to the hospital; 48 (0.48%) were discharged to hospice; 41 (0.41%) expired; 23 (0.23%) left against medical advice; and 46 (0.46%) had a durable power of attorney.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
These exclusions were applied because such patients were incapable or unlikely to complete a questionnaire (those who died and those who were discharged to the hospital), patients for whom the burden of completing a questionnaire is potentially unethical (hospice patients who are extremely sick), or patients whose answers we could not be confident were accurate or unbiased (durable power of attorney, left against medical advice). The value of excluding these includes burden on respondents and likely distortion of the results.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

No research (to date) has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS®[1] Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical or theoretical or empirical risk adjusted or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to resident characteristics and the satisfaction scores.

[1]RTI International, Harvard University, RAND Corporation. *CAHPS Instrument for Persons Residing in Nursing Homes,* Final Report to CMS, CMS Contract No. CMS-01-01176, Sept. 2003.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not Applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Analyses used to examine SDS factors include: (1) the summary score for each of the 4 CoreQ: Short Stay Discharge questionnaire items; (2) the summary score for the CoreQ: Short Stay

Discharge measure; and (3) the summary score from the CoreQ: Short Stay Discharge measure (at the facility level).

(1) Summary Score for each of the 4 CoreQ: Short Stay Discharge Questionnaire Items

The summary score for each of the 4 CoreQ: Short Stay Discharge questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. Correlation and T-test a**nalyses were** used to compare the SDS means with each other (Table 2b4.4b.a). These analyses show that the individual item scores used in the CORE Q: Short Stay Discharge measure are not significantly different based on either education level or race. That is, the educational makeup of the respondents or the racial makeup of the respondents does not influence the scores for individual items.

Table 2b4.4b.a: Mean CoreQ: Short Stay Discharge Item Distribution by Education

What is the highest grade or level of school that you have completed?	Respondents	**Q1**	**Q2**	**Q3**	**Q4**
		Mean	*Mean*	*Mean*	*Mean*
Some high school, but did not graduate	***10% (n=103)***	3.99	3.96	4.00	3.93
High school graduate or GED	***36% (n=363)***	3.83	4.03	3.99	3.85
Some college or 2 year degree	***25% (n=256)***	3.83	3.94	3.79	3.80
4 year college graduate	***17% (n=175)***	3.81	3.94	3.93	3.94
More than 4 year college degree	***11% (n=114)***	3.99	3.89	3.94	4.01
RANK CORRELATION		0.0094	0.0413	0.0418	0.0374

RANK CORRELATION OF ITEMS WITH EDUCATION: NONE SIGNIFICANT AT p=0.05

Table 2b4.4b.a (continued): Mean CoreQ: Short Stay Discharge Item Distribution by Race

What is your race?	**Respondents**	**Q1**	**Q2**	**Q3**	**Q4**
		Mean	Mean	Mean	Mean
White	***95% (n=972)***	3.87	3.99	3.94	3.89
Black or African-American	**3% (n=26)**	3.69	3.79	3.77	3.92
Asian	**2% (n=16)**	4.18	4.06	4.01	4.06
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0	0

TWO-SAMPE T-TEST	1 vs. 2	0.43	0.33	0.88	0.41
	1 vs. 3	0.27	0.78	0.54	0.5
	2 vs. 3	0.15	0.43	0.68	0.33

RACE ITEMS: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(2) Summary Score for the CoreQ: Short Stay Discharge Measure

The summary score for each of the 4 CoreQ: Short Stay Discharge questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 4 questionnaire items the average score for the resident is calculated. Correlation and T-test a**nalyses were** used to compare the SDS means with each other (Table 2b4.4b.b). These analyses show that the CORE Q: Short Stay Discharge measure score is not significantly different based on either education level or race of respondents. That is, the educational makeup of the respondents or the racial makeup of the respondents does not influence the measure score.

Table 2b4.4b.b: Mean CoreQ: Short Stay Discharge Distribution by Education

What is the highest grade or level of school that you have completed?	Respondents	Measure Score
		Mean
Some high school, but did not graduate	***10% (n=103)***	3.96
High school graduate or GED	***36% (n=363)***	3.93
Some college or 2 year degree	***25% (n=256)***	3.84
4 year college graduate	***17% (n=175)***	3.91
More than 4 year college degree	***11% (n=114)***	3.97

RANK CORELATION .0066
RANK CORRELATION OF MEASURE SCORE WITH EDUCATION: NOT SIGNIFICANT AT p=0.05

Table 2b4.4b.b (continued): Mean CoreQ: Short Stay Discharge Distribution by Race

What is your race?	Respondents	Measure Score
		Mean
White	***95% (n=972)***	3.92
Black or African-American	**3% (n=26)**	3.76
Asian	**2% (n=16)**	4.01
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0
American Indian or Alaskan Native	**0% (n=0)**	0
TWO-SAMPLE T-TEST	1 vs. 2	0.41
	1 vs. 3	0.50
	2 vs. 3	0.33

RACE MEASURE SCORE: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(3) Summary score from the CoreQ: Short Stay Discharge Measure (at the Facility Level).

The summary score for each of the 4 CoreQ: Short Stay Discharge questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 4 questionnaire items the average score for the resident is calculated. The facility score represents the percent of residents with average scores of 3 or above. A t-test a**nalysis was** used to compare the mean scores (Table 2b4.4b.c). This analysis demonstrated the CORE Q: Short Stay Discharge measure is not significantly different based on either education level or race. That is, the educational makeup of the respondents or the racial makeup of the respondents does not influence the measure.

Table 2b4.4b.c: CoreQ: Short Stay Discharge Score with and without adjustment for Education

What is the highest grade or level of school that you have completed?	Respondents	Measure Score		
		Score with SDS Characteristic vs. Without Characteristic		
Some high school, but did not graduate	***10% (n=103)***	83.4	83.2	n.s
High school graduate or GED	***36% (n=363)***	83.4	83.1	n.s
Some college or 2 year degree	***25% (n=256)***	83.4	82.9	n.s
4 year college graduate	***17% (n=175)***	83.4	83.1	n.s
More than 4 year college degree	***11% (n=114)***	83.4	83.8	n.s

N.S. = Not significant at p=0.05

Table 2b4.4b.c (continued):CoreQ: Short Stay Discharge Score with and without adjustment for Race

What is your race?	Respondents	Measure Score		
		Mean		
White	***95% (n=972)***	83.4	83.3	n.s
Black or African-American	**3% (n=26)**	83.4	83.4	n.s
Asian	**2% (n=16)**	83.4	83.4	n.s
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0

N.S. = Not significant at p=0.05

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable

2b4.9. Results of Risk Stratification Analysis:
Not Applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not Applicable

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
We performed an analysis to examine whether the CoreQ Short Stay Discharge measure captured clinically/practically meaningful differences between providers. We produced a histogram of the scores for the providers in the Pilot CoreQ: Short Stay Discharge questionnaire sample (figure 1b.2).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
See appendix for a histogram in 1b.2 (figure 1b.2) showcasing the distribution of scores.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The CoreQ Short Stay Discharge scores reflect practical and meaningful differences in quality between facilities. The histogram in Section 2b5.2 (figure 1b.2) shows that the distribution of summary scores is quite wide, indicating the scores can be used to differentiate facilities of varying levels of customer satisfaction quality.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not Applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS
2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Four items are used in the CoreQ: Short Stay Discharges questionnaire. In calculating the CoreQ: Short Stay Discharge measure if 1 item of 4 is missing then imputation is used, and if 2 (or more) of the 4 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. The testing to identify the extent and distribution of missing data included examining the frequency of missing responses for each of the 4 CoreQ: Short Stay Discharges questionnaire items and the extent and distribution of missing data for more than one missing response for the items. The method of testing to identify if the performance results were biased included examining the correlation with the quality indicators (described above) when imputation was and was not used.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
As noted above in section 2b7.1, 4 items are used in the CoreQ: Short Stay Discharges questionnaire. In calculating the CoreQ: Short Stay Discharge measure if 1 item of 4 is missing then imputation is used, and if 2 (or more) of the 4 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. From the testing of 10,319 residents (described elsewhere) we found:

1. In recommending this facility to your friends and family, how would you rate it overall?
   - That missing responses occurred in 3.71% (n=383) cases.
2. Overall, how would you rate the staff?
   - Missing responses occurred in 3.54% (n=365) cases.
3. How would you rate the care you receive?
   - Missing responses occurred in 3.9% (n=402) cases.
4. How would you rate how well your discharge needs were met?
   - Missing responses occurred in 5.21% (n=538) cases.

Two (or more) missing responses occurred in 347 cases. Thus, the degree of missing data was very small (=2.4%). Imputation was used in 1341 cases or 12.9% of respondents.

Using the cases with 1 missing value (i.e., those with imputation) the correlation with the quality indicators described above (i.e., restraint use, pressure ulcers, catheter use, antipsychotic use, antidepressant use, antianxiety use, use of hypnotics, and deficiency citations) was unchanged compared to those with no imputation.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Bias from imputation was minimal. The correlation with the quality indicators described above (i.e., restraint use, pressure ulcers, catheter use, antipsychotic use, antidepressant use, antianxiety use, use of hypnotics, and deficiency citations) was unchanged. When the respondents were removed from the analyses, the average summary scores remained the same.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Satisfaction Survey

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Since the CoreQ: Short Stay Discharge measure has been created and utilized in testing and quality improvement, we have modified it in the following ways.
Additionally, we examined how frequently facilities could administer the questionnaire and the impact of waiting longer periods. We recommend that a facility administer the questionnaire weekly (but up to 2 weeks after patient discharge). The facility operating systems are able to generate patient records after these intervals (i.e., 1 week and 2 weeks). Furthermore, it is advantageous if administered weekly as we identified an increase in response rate of approximately 8%. Moreover, this time period is optimal in order to minimize recall bias. Therefore, this recommendation was incorporated into the measure specifications (given above).

We conducted analyses on allowing up to 2 months for a patient to respond. We identified the average (modal) response to occur within 2 weeks. A few responses were still received 6 weeks after administration, however, by 2 months the response was very much lower (<5% of additional returned surveys). Furthermore, in order to ensure that this time frame did not bias the type of responses captured, we analyzed the average score for the surveys returned. We found that the average scores for surveys returned in the first month were almost identical to those returned in the second month. Thus, this recommendation was incorporated into the measure specifications (given above).

We examined the effect of the 6 month survey completion time period on a facility's ability to collect the survey data. Even the largest nursing facilities need an extended period of time to achieve the 20 minimum sample size identified above. We identified that a majority of nursing facilities (i.e., 90%) in our sample could achieve this response rate if given up to 6 months. Therefore, this recommendation was incorporated into the recommendations (given above).
We conducted analyses on collecting data from residents discharged to the hospital. We identified that patients discharged to the hospital did not have high response rates (i.e., 1 out of 25 were returned). Therefore, discharge to an acute care hospital became an exclusion criterion.

Furthermore, we decided that once 125 consecutive responses are received for a particular facility, it is optional to stop the collection prior to the 6 month period and calculate the measure, because past this mark, no additional information effects the SNFs satisfaction score. Moreover, at 125 responses, the confidence interval shrinks, increasing the certainty of the CoreQ: Short Stay Discharge questionnaire as capturing the true population customer satisfaction.

As part of the CoreQ: Short Stay Discharge measure development, existing satisfaction vendors were contacted (including MyInnerView, Symbria, and NRC) for input on the administration and sample selection used. With respect to administration, the 2 month window used for including returned surveys and the 2 week period from discharge to administer the survey were viewed positively and are currently standard time periods used in the industry. With respect to the sample selection, the exclusion criteria (i.e., Patients who die; patients who were discharged to a hospital, another SNCC, psychiatric facility, Inpatient Rehabilitation Facility, or MR/DD facility; patients with Court appointed legal guardian for all decisions; patients on hospice; patients who left the nursing facility against medical advice) were well received by these vendors. In many cases most of these sample selection criteria are already used by the vendors. Also, with respect to the sample selection, the use of the MDS to capture the sample selection criteria (above) were well received by these vendors.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm) exist.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) AHCA Quality Initiative https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx Massachusetts Senior Care N/A Satisfaction Vendors N/A Quality Improvement (Internal to the specific organization) Large Nursing Home Chain N/A

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The CoreQ: Short Stay Discharge measure is currently in use by a large nursing home chain for the purposes of quality improvement.

In addition, Massachusetts Senior Care is currently using the Measure for quality improvement. A total of 150 facilities in Massachusetts are collecting satisfaction data using of the CoreQ: Short Stay Discharge questionnaire. The CoreQ: Short Stay Discharge measure will be calculated and distributed in a report card to each participant (this is currently on-going).

Furthermore, 10 large national satisfaction vendors in the SNF area have agreed to add the CoreQ to their questionnaires and calculate the measure. The following customer satisfaction vendors are using CoreQ.

- Align
- Brighton Consulting Group
- Healthcare Academy (ReadyQ)
- inQ Experience Surveys
- National Research Corporation (My Innerview)
- Pinnacle
- Providigm/abaqis
- Sensight Surveys
- Service Trac
- The Jackson Group, Inc.

We do not have counts of patients being surveyed and geographical representation from the vendors, however they represent the majority of customer satisfaction vendors currently doing SNF business in the United States.

A letter has been sent to all 10,000 AHCA SNF members indicating which vendors to date have agreed to add the CoreQ to their questionnaire and calculate the measure (see attached letter in appendix, section 4.a.1). A user's manual has been developed and is available on AHCA's website for all satisfaction survey vendors to use.

AHCA and NCAL have also incorporated the CoreQ into their national Quality Initiative goals. AHCA represents nearly 10,000 of the 15,000 SNFs and provides feedback to all of its members on their satisfaction scores using the CoreQ. This has resulted in growing number of members and vendors collecting the data.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The CoreQ: Short Stay Discharge measure is not currently publicly reported or used in other accountability applications (e.g., payment program, certification, licensing). The reason for this is that it is a new measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
AHCA has recently started the second Quality Initiative, laying out a series of quality improvement and reporting goals for the AHCA membership, which covers over 9,600 of all SNFs in the U.S. Among these goals is the collection and reporting of CoreQ customer satisfaction data. Because it has been included in the Quality Initiative 2015-2018, AHCA's machinery for publicizing and encouraging the adoption of the tool has been activated, including AHCA's quality division spending a large number of staff hours working to accomplish this. In addition to marketing the use of the survey instrument as a way for SNFs to understand how their patients view the care and other services that they were provided by the SNFs, AHCA is developing an upload and reporting feature within its member data profiling tool, LTC Trend TrackerSM, which allows SNFs to centrally view a large number of quality, compliance, operational and financial metrics from public and non-public sources. The CoreQ report and upload feature within LTC Trend Tracker will include an API for vendors performing the survey on behalf of SNFs – AHCA's preferred approach to collecting the data – so that the aggregate CoreQ results will be immediately available to providers as they are collected. Given that LTC Trend TrackerSM is probably the leading method for SNFs to profile their quality and other data, the incorporation of CoreQ into LTC Trend Tracker means it will immediately become the de facto standard for customer satisfaction surveys for the SNF industry.

In addition, large national satisfaction vendors in the SNF area have agreed to add the CoreQ to their questionnaires and calculate the measure. An email letter has been sent to all 10,000 AHCA SNF members indicating which vendors to date have agreed to add the CoreQ to their questionnaire and calculate the measure (see attached letter in appendix section 4a.1).

We also are working with states who require satisfaction measurement to incorporate the CoreQ into their process. The State of Rhode Island pilot tested a version of the CoreQ in its statewide satisfaction questionnaire for long stay residents. The state of Massachusetts has included the CoreQ short stay as part of its current ongoing deliberation on measuring satisfaction in SNFs. AHCA has a presence in each state, and our state affiliates will be promoting the use of the CoreQ in those states that are collecting or considering collecting satisfaction.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not Applicable

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no negative consequences to individuals or populations identified during testing or evidence of unintended negative consequences to individuals or populations reported since the implementation of the CoreQ: Short Stay Discharge questionnaire or the measure that is calculated using this questionnaire.

This is consistent with satisfaction surveys in general in nursing facilities. Many other satisfaction surveys are used in nursing facilities with no reported unintended consequences to patients or their families.

There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make those patients further dissatisfied.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** CoreQ_Short_Stay_Appendix_Final.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Health Care Association
Co.2 **Point of Contact:** Urvi, Patel, upatel@ahca.org, 202-898-2858-
Co.3 **Measure Developer if different from Measure Steward:** American Health Care Association
Co.4 **Point of Contact:** Lindsay, Shwartz, lshwartz@ncal.org, 202-898-2848-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The workgroup gave input, reviewing our suggested administration, required response rate, the manual, and exclusions.

Mary Tess Crotty, Genesis - Also helped provide feedback on the development process and the user manual. Additionally, she reviewed the analyses.
Matt O'Connor HCR Manor Care- Also helped provide feedback on the development process and the user manual. Additionally, he conducted some analyses and reviewed the analyses.
Judy Hoff, Health Care Academy
Rich Kortum, My Innerview/National Research Corporation
Peter Kramer, abaqis/Providigm
Ellen Kuebrich, abaqis/Providigm
Michael Johnson, ServiceTrac
Chris Magelby, Pinnacle

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:** 10, 2015

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.6 **Copyright statement:** None

Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2615
De.2. **Measure Title:** CoreQ: Long-Stay Resident Measure
Co.1.1. **Measure Steward:** American Health Care Association
De.3. **Brief Description of Measure:** The measure calculates the percentage of long-stay residents, those living in the facility for 100 days or more, who are satisfied (see: S.5 for details of the time-frame). This patient reported outcome measure is based on the CoreQ: Long-Stay Resident questionnaire that is a three item questionnaire.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.

The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.
As part of the US Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement

system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay questionnaire and measure can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007). This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Long Stay questionnaire and measure are needed to improve the care for long stay SNF patients.

Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long Stay measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

Numerator Statement: The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long -Stay Resident questionnaire.
Denominator Statement: The denominator includes all of the residents that have been in the SNF for 100 days or more regardless of payer status; who received the CoreQ: Long-Stay Resident questionnaire (e.g. people meeting exclusions do not receive the questionnaire), who responded to the questionnaire within the two month time window, who did not have the questionnaire completed by somebody other than the resident, and who did not have more than one item missing.
Denominator Exclusions: Exclusions made at the time of sample selection are the following: (1) Residents who have poor cognition defined by the BIMS score; (2) residents receiving hospice; (3) residents with a legal court appointed guardian; and (4) residents who have lived in the SNF for less than 100 days.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (two months after the administration date) b) surveys that have more than one questionnaire item missing c) surveys from residents who indicate that someone else answered the questions for the resident. (Note this does not include cases where the resident solely had help such as reading the questions or writing down their responses.)

Measure Type: PRO
Data Source: Healthcare Provider Survey
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This is a patient-reported outcome measure of patient satisfaction for patients in skilled nursing facilities (SNFs). The developer provides a diagram and a table demonstrating the links between structures and/or processes and the outcomes influence patient satisfaction, and the final patient reported outcome of satisfaction.
- The developer notes that “Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management”

- The developer states "We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care" and notes that measuring patient satisfaction is required for person-centered care for three reasons:
    - Measuring satisfaction is necessary to understand patient preferences.
    - Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
    - Satisfaction information can help facilities improve the quality of care they provide
- Finally, the developer states "The definition of quality in a nursing facility has shifted from a focus on structure and process criteria to clinical outcomes, resident satisfaction, and quality of life".

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes.
- Studies show a link between patient satisfaction and the following health-related behaviors:
    - Keeping follow-up appointments
    - Disenrollment from health plans
    - Litigation against providers

The developer provided performance scores based on 194 facilities that met the inclusion criteria (20 valid responses and 30% response rate). The scores include tables by age and gender. The facility score represents the percent of residents with average scores of 3 (good) or above (scale 1-5).

Table 1b.2.c: Facility Level Performance Distribution

Variable	Observations	Mean	Standard Deviation	Minimum	Maximum
CoreQ2	194	3.628701	.3413973	2.55814	4.842105
CoreQ3	194	3.58848	.3584188	2.454545	4.789474
CoreQ1	194	3.527274	.3892386	2.272727	4.842105

Table 1b.2.d: Overall Descriptive Information for the Summary Score MEASURE

	Minimum	p25	p50	p75	Maximum
Summary Score	35.6	59.0	64.7	70.0	95.6

Disparities

The developer says differences in scores based on SDS categories was not statistically significant:

- By race/ethnicity, whites averaged a score of 83.2, Blacks or African-Americans averaged a score of 83.3, and Asians 83.4
- By highest education level those with those high school but who did not graduate averaged 83.2, high school graduates averaged 83.5, those with some college or a 2-year degree averaged 82.5, 4 year college graduates averaged 83.4, and those with more than 4 year college degree averaged 83.8
- By age group, residents younger than 65 years old averaged 72.9, those 65-74 averaged 82.7, those 75-84 averaged 85.0, and those older than 85 averaged 85.0.
- by gender, males averaged a score of 81.1 and females averaged a score of 83.9.

However, research over the last 20 years has consistently found poorer care in facilities with high minority populations and that nursing homes remain segregated, with black patients concentrated in poorer-quality homes (as measured by staffing ratios, performance, and are more financially vulnerable).

The developer did not risk adjust this measure for SES because "adjusting for racial status has the unintended effect of adjusting for poor quality providers not to differences due to racial status and not within-provider discrimination."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**Patient experience measures are not standardized and there is a call for more experience data in long term care.

**The developer provides several references indicating that better performance on several process and structural measures of care quality are associated with higher satisfaction. However, while they do not provide it, there are data supporting that patient satisfaction and outcomes do not always correlate closely. However, based upon the national focus on PROs and the existence of other NQF-endorsed satisfaction measures, I agree with NQF staff that this measure passes this criterion.

**Measure responds to evidence of two nh populations, short and long-term. With long-term pop having lower satisfaction rates than those who return home. Not surprising but response to assess reason for lt pop dissatisfaction appropriate. Lack of standardized measures also called for this approach. Recent legislative action calls for QI and accountability. Drivers of satisfaction are identified and linked to the patient-reported outcome, primarily linked to staff.

Have concerns about the potential large number of exclusions from the denominator. I have trouble seeing the value of this broad measure.

**Not surprising but response to assess reason for lt pop dissatisfaction appropriate. Lack of standardized measures also called for this approach. Recent legislative action calls for QI and accountability. Drivers of satisfaction are identified and linked to the patient-reported outcome, primarily linked to staff. Have concerns about the potential large number of exclusions from the denominator. I have trouble seeing the value of this broad measure.

1b. Performance Gap

Comments:

**There is variability in results so opportunity is present.

**The developer demonstrates a 10 absolute percentage point difference between Q25 and Q75 in the proportion of patients within a facility achieving average good or better score across the CoreQ (59% to 70%). I agree with NQF staff that this measure meets a moderate response for this this criterion.

**Disparities between subgroups was not significant and no risk adjustment was made.

1c. Performance Gap

Comments:

**Focus groups of families and patients were involved.

**The developer sought input from a focus group of 40 patients who noted that the CoreQ questions/domains were important to them. I believe this satisfies a moderate response for this criterion.

**I do not see how this case was made, but believe that snf residents value quality care, staffing levels and appropriate training. Patient satisfaction linked to lower rates of litigation also.

Criteria 2: Scientific Acceptability of Measure Properties

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): CoreQ: Long Stay Resident questionnaire – health care provider

Specifications:

- The level of analysis is facility.
- The measure result is a non-weighted percentage score:
  - the sum of the individuals in the facility that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long -Stay Resident questionnaire over all of the residents that have been in the SNF for 100 days or more regardless of payer status; who received the CoreQ: Long-Stay Resident questionnaire (e.g. people meeting exclusions do not receive the questionnaire), who responded to the questionnaire within the two month time window, who did not have the questionnaire completed by somebody other than the resident, and who did not have more than one item missing.
- There is no data dictionary.
- A calculation algorithm is described.
- **The measure is not risk adjusted or stratified.**
- There are exclusions.
- The following are excluded from the sample:
  - Residents who have poor cognition defined by the BIMS score;
  - residents receiving hospice;
  - residents with a legal court appointed guardian; and
  - residents who have lived in the SNF for less than 100 days.
- Once the survey is administered, the following are excluded from the responses:
  - surveys received outside of the time window (two months after the administration date)
  - surveys that have more than one questionnaire item missing
  - surveys from residents who indicate that someone else answered the questions for the resident. (Note this does not include cases where the resident solely had help such as reading the questions or writing down their responses.)

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*

○ *Is the logic or calculation algorithm clear?*

○ *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing [Testing attachment]

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data elements were tested using a test-retest methodology. The pilot survey included 100 residents and the follow-up survey included 50 residents. The distribution of responses and the correlation between the original and follow-up scores were then calculated.
- Person/questionnaire level was tested using the same test-retest methodology.
- The stability of the facility-level score was tested using bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score.

Results of reliability testing

Results for each level of testing are presented.

Average Percent Agreement between the Pilot and Re-administered Survey – Data Element testing

Questionnaire Item	Percent Agreement
1. In recommending this facility to your friends and family, how would you rate it overall?	97.6%
2. Overall, how would you rate the staff?	98.5%
3. How would you rate the care you receive?	98.0%

Average Percent Agreement between Response Options for the Pilot Survey and Re-Administered Survey - Person/Questionnaire level testing

		Re-Administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
Pilot Response	Poor (1) or Average (2)	98.75%	98.5%
	Good (3), Very Good (4), or Excellent (5)	98.75%	99%

Facility level testing:

In the 10,000-repetition bootstrap, scores are moderately stable; the minimum sample size was 20 and the maximum 122.

- 14.18% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample,
- 20.91% were within 3 percentage points,
- 33.50% were within 5 percentage points, and
- 46.33% were within 10 percentage points.

Guidance from the Reliability Algorithm
Submitted specifications precise and complete (Box 1): Yes →Empirical reliability testing conducted (Box 2): Yes →Reliability testing conducted with computed performance measure score (Box 4): Yes →Method described appropriate (Box 5): Yes → Level of certainty/confidence that the performance measure scores are reliable (Box 6) → Moderate (based on developer assessment of moderate stability)

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☒ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
1. Validity testing of the questionnaire format used in the CoreQ: Long-Stay Resident questionnaire
    - Face validity evaluated via literature review and review of 12 commonly used satisfaction surveys; also examined face validity of domains and the response scale, using 40 patients in 5 nursing homes. The Flesch-Kinkaid scale was used to determine if patients understood the questions.

2. Testing the items for the CoreQ: Long-Stay Resident questionnaire;
    - Exploratory factor analysis (EFA) were used to further refine the pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items.

1. To determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Long Stay Resident measure);
    - Correlation analysis and a factor analysis conducted on items using Cronbach's Alpha

4. Validity Testing for the CoreQ: Long-Stay Resident measure.

- Developers examined correlation between the three items in the measure and all of the items on the pilot instrument.
- Also examined correlations between the CoreQ: Long-Stay Resident measure scores and measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, and several other quality metrics from Nursing Home Compare

Validity testing results:

Results for each level of validity testing are provided.

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Resident Questionnaire
    A. The literature review shows that domains used in the Pilot CoreQ: Long-Stay Resident questionnaire items have a high degree of both face validity and content validity.
    B. Residents overall rankings, show the general "domain" areas used indicates a high degree of both face validity and content validity.
    C. The results show that 100% of residents are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.
    D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the item.

2. Testing the Items for the CoreQ: Long-Stay Resident Questionnaire
    A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks etc.
    B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

3. Testing to Determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Resident measure)

    A. Using the correlation information of the Core Q: Long-Stay Resident questionnaire (18 items) and the 3 items representing the CoreQ: Long-Stay Resident questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.

    B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Long-Stay Resident Measure

    A. The correlation of the 3 item CoreQ: Long-Stay Resident measure summary score (with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.89. This indicates that the CoreQ: Long-Stay Resident measure score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

    B. Relationship with CASPER Quality Indicators: The 8 CASPER Quality Indicators all had a reasonable level of negative correlation with the CoreQ: Long-Stay Resident measure in the direction as expected (higher satisfaction is associated with better quality. These correlations range from -0.105 to -0.476. The CoreQ: Long-Stay Resident measure is associated with these quality indicators. This testing indicates a reasonable degree of construct validity and convergent validity.

C. Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels

- The 13 Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels all had a moderate to high level of correlation and in the direction predicted with the CoreQ: Long-Stay Resident measure. These correlations range from ± 0.100 to 0.47. The CoreQ: Long-Stay Resident measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.47 for RN hours per resident-day and 0.37 for total staffing hours per resident day. This testing indicates a reasonable degree of construct validity and convergent validity.
- As noted by Mor and associates (2003, p.41) "there is only a low level of correlation among the various measures of quality" In long term care settings. Castle and Ferguson (2010) also show the pattern of findings of quality indicators in nursing facilities is consistently moderate with respect to the correlations identified. The magnitude of correlations of the CoreQ with quality metrics are consistent with these findings in this setting.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- An expert panel advised the developer on exclusions. They were advised to 1) Residents with dementia impairing their ability to answer the questionnaire defined as having a low BIMS score; (2) residents receiving hospice care; and (3) Residents with a legal court appointed guardian. In addition, the developer elected to exclude (4) Residents who have lived in the SNF for less than 100 days; (5) Respondents who have one or more missing data point (on the COREQ items); and (6) residents without usable data defined as missing data on 2 or 3 of the 3 questions.
- The developer reports that these are commonly used with satisfaction surveys and that since "the exclusions were based on individual's ability to answer questions and were also made in the pilot, we are not able to confirm if the exclusions actually made a difference to the scores, which is why we cannot calculate the mean CoreQ: Long-Stay Resident scores with and without the exclusions". They further note that residents were excluded because they were unable to provide an independent response, the burden was inappropriate, or the developer could not be confident in the answers. "Therefore, the value of excluding these residents takes into account burden on respondents and their ability to answer the questions."
- The exclusion analysis included 223 facilities and included 34% of residents who have poor cognition; 2% residents with hospice; and 4% residents with a legal court appointed guardian.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

The developer stated: No research (to date) has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS® Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical, theoretical, or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to resident characteristics and the satisfaction scores.

Questions for the Committee:

- *If a justification for no risk adjustment is provided, is there any evidence that contradicts the developer's rationale and analysis?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states that "the CoreQ Long-Stay Resident scores reflect practical and meaningful differences in quality between facilities."
- A histogram is provided.

Question for the Committee:

Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- Three items are used to calculate the measure. If one item is missing, imputation is used to calculate. If two or more items are missing, the respondent is excluded. Two (or more) missing responses occurred in 123 cases out of 7,307. The degree of missing data was 1.68%.
- The imputation method consists of using the average score from the items answered. Imputation was used in 904 cases or 12.37% of respondents.
- The developer states that "Bias from imputation was minimal due to the rate of missingness being very low. … When the respondents were removed from the analyses, the average Summary Scores remained the same."

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm:
Measure specifications consistent with evidence (Box 1): Yes →Potential threats to validity empirically assessed (Box 2): Yes →Empirical Validity testing conducted (Box 3): Yes (note, face validity also assessed) →Validity testing conducted with computed measure score (Box 6): Yes →Method(s) described appropriate (Box 7): Yes →Level of certainty or confidence that performance measure scores are a valid indicator of quality (Box 8): Moderate

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b)

2a.1 Specifications

Comments:

**Survey also compared to other existing data sets.
**The specifications are precise and clear. As above, they confirmed that the relevant patient population finds the component questions meaningful and important.
**The measure seems to be the broadest possible addressing general satisfaction at a given point when the survey is taken. I am not clear what other data elements there are. it could be consistently implemented. But still do not like exclusions.

2a.2 Reliability Testing

Comments:

**Testing occurred at patient level and facility level for reliability with consistency of results.
**Data element reliability was tested across a combined sample of 150 patients using a test-retest approach and measure result reliability was tested using a bootstrap method using data from 223 facilities. While I would have liked to see the facility-level measure result reliability testing using the test-retest approach, I agree with NQF staff that this measure meets a moderate response for this this criterion.

**Reliability relatively high.

2b.2 Validity Testing

Comments:

**The strongest evidence for validity are the data demonstrating that the measure results are correlated with clinical outcomes. The developers provide extensive evidence of the validity of the PROM instrument itself, however they only demonstrate concurrent validity between the CoreQ and its parent survey which does not strike me as sufficient evidence of validity.

My only concern about validity is that the PROM score for the CoreQ uses a 5 point Likert scale for the response that places "good" as the center item (poor, average, good, very good, excellent). I am unfamiliar with the validity of using a Likert scale where the centroid is not neutral; this strikes me as an important threat to validity that is not addressed in the application.

I defer rating this criterion to allow a larger discussion with the committee and NQF staff.

2b.3-2b7. Testing

Comments:

**Exclusions are present if the resident didn't fill it out themselves. Given patient and family centeredness as a goal, would the facilities want the family input too if the resident can't provide it?

**This measure is not risk adjusted but other satisfaction surveys (CAHPS) are risk adjusted to account for language/other factors. Also, although they found no significant disparities, the data sample was not very diverse or representative to allow adequate power to detect these differences.

Missing data were uncommon and facility level results were unchanged with missing data included.

I defer rating this criterion to allow a larger discussion with the committee and NQF staff.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This is a patient satisfaction survey. The developer states that 90% of facilities can meet the sample size of 20 surveys within the two-month period allotted.
- No fees required to use the measure; survey is available via AHCA

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**86% of the facilities had a response rate over 30% / No comment made on whether the survey is available in other languages

**As a PRO-PM, this requires additional data collection and the developer notes the PROM doesn't require licensing fees. However, the PROM used seems to either require use of an existing vendor or can be collected by the developer "at a discounted price". This suggests some cost to facilities but this cost may be incrementally insignificant - I would appreciate some clarity form either NQF staff or developer on the true cost to facilities to implement this measure as that is not clear.

I defer rating this criterion to allow a larger discussion with the committee and NQF staff.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

- AHCA Quality Initiative: https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx
- Satisfaction Vendors (10 national companies)

Quality Improvement (Internal to the specific organization)

- Large Nursing Home Chain

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details
Not in use for accountability program, but ACHA plans to begin public reporting of the CoreQ measures as part of their Quality Initiative 2016-2018 (nearly 10,000 of 15,000+ Medicare & Medicaid certified SNFs)

Improvement results
New measure – no information available

Unexpected findings (positive or negative) during implementation
New measure – no information available

Potential harms
The developer states, "There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make those patients further dissatisfied."

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**Plan for use in accountability.

**ACHA plans to include the CoreQ PROM in its Quality Initiative in 2016-2018 for CMS-certified SNFs.

The skewing of the Likert scale responses may under-report poor performance, leading to an under-appreciation or poor care.

If the issues regarding the Likert scale and risk adjustment can achieve consensus within the committee, I think this measure meets moderate usability criterion.

Criterion 5: Related and Competing Measures

Related or competing measures

None – Measure 0692 listed in the submission is no longer NQF-endorsed

Harmonization

N/A

Pre-meeting public and member comments

-

Submission materials attachments…

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: *CoreQ: Long Stay Resident Measure*

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: **Instructions**

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the

strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☒Patient-reported outcome (PRO): Customer Satisfaction

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

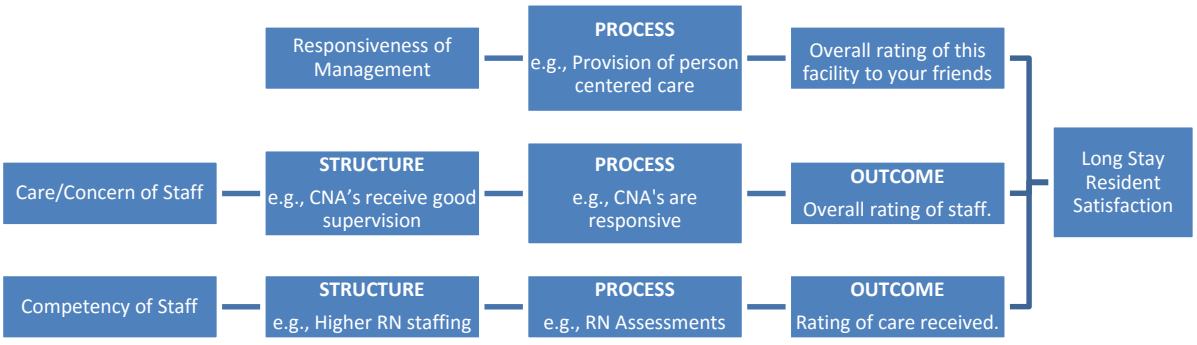
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Long stay satisfaction can be looked at as the outcome for a number of structures and processes within skilled nursing care centers. Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management (National Research Corporation, 2014).



Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the*

Health Professions, 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

National Research Corporation. (2014). 2014 National Research Report Empowering Customer-Centric Healthcare Across the Continuum.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

The table below provides the structure and process drivers that influence long stay resident satisfaction.

Authors	***Structure or Process and Driver of Long Stay Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Reinhardt, et al., 2014	Process Care/concern of staff and competency of staff	Conversations regarding end-of-life care options with family members show higher overall satisfaction with care and more use of advance directives.	Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. *Journal Of Social Work In End-Of-Life & Palliative Care*. 10(2):112-26.
Van Uden et al. (2013).	Process Competency of staff	For nursing home residents with dementia improved symptom management is associated with higher satisfaction with care.	van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. *International Psychogeriatrics.* 25(10):1697-707.
Li et al. (2013).	Structure Responsiveness of management	Higher overall nursing home satisfaction scores were associated with higher nursing staffing levels and	Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care was generally high during 2005-09, with some variability

		fewer deficiency citations.	across facilities. *Health Affairs*. 32(8):1416-25.
Crogan et al. (2013).	Structure Responsiveness of management	Improvements in a nursing home food delivery system were associated with higher overall satisfaction and improved resident health.	Crogan, N.L., Dupler, A.E., Short, R., & Heaton, G. (2013). Food choice can improve nursing home resident meal service satisfaction and nutritional status. *Journal of Gerontological Nursing*. 39(5):38-45.
Authors	***Structure or Process and Driver of Long Stay Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Brownie & Nancarrow (2013).	Structure & Process Responsiveness of management and care/concern of staff	Implementation of person-centered care is associated with higher levels of satisfaction.	Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. *Clinical Interventions In Aging*. 8:1-10.
Kleijer et al., 2014	Process Competency of staff	Residents perceive a low level of quality of care in centers where there is a high level of antipsychotic use.	Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts, A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.
Kayser-Jones et al., 1999	Structure Responsiveness of management and	Higher levels of RN and LPN staffing have been associated with better quality outcomes such as ADL maintenance and hydration. Centers that	Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision.

	care/concern of staff	have a family council in addition to the required resident council have higher resident satisfaction.	*Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the Health Professions,* 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts, A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.

Bishop, C., Weinberg, D., Leutz, W., Dossa, A., Pfefferle, S., & Zincavage, R. (2008). Nursing assistants' job commitment: Effect of nursing home organizational factors and impact on resident well-being. *The Gerontologist*, 48(1), 36-45.

Lucas, J.A., Lowe, T.J., Robertson, B., Akincigil, A., Sambamoorthi, Q., Bilder, S., Paek, E.K., & Crystal, S. (2007). The relationship between organizational factors and resident satisfaction with nursing home care and life. *Journal of Aging & Social Policy,* 19(2), 125-151.

Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision. *Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Kane, R.L., & Kane, R.A. (2001). What older people want from long-term care, and how can they get it. *Health Affairs*, 20(6), 114-127.

Westat. Resident experience with nursing home care: A literature review.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2615
De.2. **Measure Title:** CoreQ: Long-Stay Resident Measure
Co.1.1. **Measure Steward:** American Health Care Association
De.3. **Brief Description of Measure:** The measure calculates the percentage of long-stay residents, those living in the facility for 100 days or more, who are satisfied (see: S.5 for details of the time-frame). This patient reported outcome measure is based on the CoreQ: Long-Stay Resident questionnaire that is a three item questionnaire.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.

The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.
As part of the US Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay questionnaire and measure can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007).

This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Long Stay questionnaire and measure are needed to improve the care for long stay SNF patients.

Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long Stay measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

S.4. **Numerator Statement:** The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long -Stay Resident questionnaire.
S.7. **Denominator Statement:** The denominator includes all of the residents that have been in the SNF for 100 days or more regardless of payer status; who received the CoreQ: Long-Stay Resident questionnaire (e.g. people meeting exclusions do not

receive the questionnaire), who responded to the questionnaire within the two month time window, who did not have the questionnaire completed by somebody other than the resident, and who did not have more than one item missing.
S.10. **Denominator Exclusions:** Exclusions made at the time of sample selection are the following: (1) Residents who have poor cognition defined by the BIMS score; (2) residents receiving hospice; (3) residents with a legal court appointed guardian; and (4) residents who have lived in the SNF for less than 100 days.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (two months after the administration date) b) surveys that have more than one questionnaire item missing c) surveys from residents who indicate that someone else answered the questions for the resident. (Note this does not include cases where the resident solely had help such as reading the questions or writing down their responses.)

De.1. **Measure Type:** PRO
S.23. **Data Source:** Healthcare Provider Survey
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not Applicable

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
CoreQ_Long_Stay__Evidence_Final.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through

their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.

The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.
As part of the US Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay questionnaire and measure can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Over the past several decades, care in nursing facilities has changed substantially. Statistics show that more than half of all elders cared for in nursing homes are now discharged home (approximately 1.6 million residents; CMS, 2009). Moreover, when satisfaction information from current residents (i.e., long stay residents) is compared with those of elders discharged home, substantial differences exist (Castle, 2007). This indicates that long stay and short stay residents are different populations with different needs in the nursing facilities. Thus, the CoreQ: Long Stay questionnaire and measure are needed to improve the care for long stay SNF patients.

Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long Stay measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The data source included 223 nursing facilities from multiple states across the US. The data were collected from June 2014 through September 2014 and included responses from 7,307 patients. The performance measure scores are available in the appendix, section 1b.2. This shows, on the 0 – 100 scale used for the CoreQ: Long-Stay Resident measure (expressed in percent), the minimum score is 35.6, the 25th percentile is 59, the 50th percentile is 64.7 the 75th percentile is 70 and the maximum score is 95.6.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We did not risk adjust the measure by sociodemographic status due to no statistically significant differences (at the 5% level) in the scores between the SDS categories. See Table 2b4.4b.b in the Testing section. By race, Whites averaged a score of 83.2, Blacks 83.3 and Asians 83.4; there were no observations for Native Hawaiians or other Pacific Islanders, American Indian or Alaskan Natives (Table 2b4.4b.c in the Testing section). By highest education level, those with some high school but who did not graduate averaged 83.2, high school graduates averaged 83.5, those with some college or a 2-year degree averaged 82.5, those with a 4-year college degree averaged 83.4, and those with more than a 4-year college degree averaged 83.3 (Table 2b4.4b.c in the Testing section). By age group, residents younger than 65 years old averaged 72.9, those 65-74 averaged 82.7, those 75-84 averaged 85.0, and those older than 85 averaged 85.0 (Table 1b.4.a in the Appendix). Furthermore, by gender, males averaged 81.1 and females averaged 83.9 (Table 1b.4.b in the Appendix).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Multiple studies in the past twenty years have examined racial disparities in the care of nursing facility residents and have consistently found poorer care in facilities with high minority populations (Fennell et al., 2000; Mor et al., 2004; Smith et al., 2007). Work on racial disparities in nursing facilities' quality of care between elderly white and black residents within nursing facility has shown clearly that nursing homes remain relatively segregated and that specifically nursing home care can be described as a tiered system in which Blacks are concentrated in marginal-quality homes (Li, Ye, Glance & Temkin-Greener, 2014; Fennell, Feng, Clark & Mor, 2010; Li, Yin, Cai, Temkin-Greener, Mukamel, 2011; Chisholm, Weech-Maldonado, Laberge, Lin, & Hyer, 2013; Mor et al., 2004; Smith et al., 2007). Such homes tend to have serious deficiencies in staffing ratios, performance, and are more financially vulnerable (Smith et al, 2007; Chisholm et al., 2013). Based on a review of the nursing facility disparities

literature, Konetzka and Werner concluded that disparities in care are likely related to this racial and socioeconomic segregation as opposed to within-provider discrimination (Konetzka and Werner 2009). This conclusion is supported, for example, by Grunier and colleagues who found that as the proportion of black residents in the nursing home increased the risk of hospitalization among all residents, regardless of race, also increased (Grunier et al., 2008). Thus, adjusting for racial status has the unintended effect of adjusting for poor quality providers not to differences due to racial status and not within-provider discrimination.

Lower satisfaction scores also likely increase as the proportion of black residents increases, indicating that the best measure of racial disparities in satisfaction rates is one that measures scores at the facility level. That is, ethnic and social economic status differences are related to inter-facility differences not to intra-facility differences in care. Therefore, the literature suggests that racial status should not be risk adjusted otherwise one is adjusting for the poor quality of the SNFs rather than differences due to racial status.

Chisholm L, Weech-Maldonado R, Laberge A, Lin FC, Hyer K. (2013). Nursing home quality and financial performance: does the racial composition of residents matter? Health Serv Res;48(6 Pt 1):2060–2080.

Fennell ML, Feng Z, Clark MA, Mor V. (2010). Elderly Hispanics more likely to reside in poor-quality nursing homes. Health Aff (Millwood);29(1):65–73.

Grabowski, D.C. (2004). The admission of Blacks to high-deficiency nursing homes. Medical Care 42(5): 456-464.

Gruneir, A., Miller, S. C., Feng, Z., Intrator, O., & Mor, V. (2008). Relationship between state Medicaid policies, nursing home racial composition, and the risk of hospitalization for black and white residents. Health Services Research, 43(3), 869-881.

Konetzka, R. T., & Werner, R. M. (2009). Review: Disparities in long-term care building equity into market-based reforms. Medical Care Research and Review, 66(5), 491-521.

Li Y, Yin J, Cai X, Temkin-Greener J, Mukamel DB. (2011). Association of race and sites of care with pressure ulcers in high-risk nursing home residents. JAMA;306(2):179–186.

Li Y, Ye Zhiqiu, Glance, Laurent & Temkin-Greener, Helena. (2014). Trends in family rating experience with care and racial disparities among Maryland nursing homes. Med Care, 52(7): 641-648.

Mor, V., Zinn, J., Angelelli, J., Teno, J. M., & Miller, S. C. (2004). Driven to tiers: socioeconomic and racial disparities in the quality of nursing home care. Milbank Quarterly, 82(2), 227-256.

Smith, D. B., Feng, Z., Fennell, M. L., Zinn, J. S., & Mor, V. (2007). Separate and unequal: racial segregation and disparities in quality across US nursing homes. Health Affairs, 26(5): 1448-1458.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, Patient/societal consequences of poor quality

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The definition of quality in a nursing facility has shifted from a focus on structure and process criteria to clinical outcomes, resident satisfaction, and quality of life. This shift was first supported by nursing home reform legislation included in the Omnibus Budget Reconciliation Act of 1987 (OBRA, 1987). Furthering the movement, the Institute of Medicine (IOM) put the patient as central to the care system (Castle, 2007; IOM, 2001) – necessitating the collection of satisfaction information. As mentioned previously (see 1b.1), a focus on person-centered care and satisfaction is also evident in the Quality Assurance & Performance

Improvement Program (QAPI) for nursing facilities and proposed Reform Requirements for Long-Term Care Facilities (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015).

Measuring and reporting satisfaction of nursing home care is important in many ways. First, residents are more likely to follow medical advice when they rate their care as satisfactory (Hall, Milburn, Roter, & Daltroy, 1998). Second, because resident satisfaction can influence the quality of care provided and the outcomes of treatment (Hudak and Wright 2000), satisfaction surveys can be used as measures of clinical and organizational accountability. Third, measuring and reporting resident satisfaction can help nursing facilities identify and improve aspects of quality. Furthermore, if publicly released, information on satisfaction with care can help elders and their families choose a nursing facility.

Several research efforts have concluded consumer satisfaction is an important indicator of quality of care in nursing homes (Bangerter et al. 2016; Shippee et al 2015; Kajonius and Kazemi, 2016; Gesell, 2001). In addition, other studies have concluded nursing resident satisfaction data provides information about quality of care that is different from clinician perspectives and clinical indicators (Berlowitz, Du, Kazis, & Lewis, 1993; Riccio 2000; Uman & Urman, 1997). This exemplifies the need for resident satisfaction data to achieve person-centered care. Only by hearing from the patient can we ensure the care provided is person-centered.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Bangerter, L.R., Heid, A.R., Abbott, K, & Van Haitsma, K. (2016). Honoring the Everyday Preferences of Nursing Home Residents: Perceived Choice and Satisfaction with Care. The Gerontologist. (Advance online publication): 1-8.

Berlowitz, D. R., Du, W., Kazis, L., & Lewis, S. (1995). Health-related quality of life of nursing home residents: Difference in patient and provider perceptions. Journal of the American Geriatric Society, 43, 799-802.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Gesell, S.B. (2001). A measure of satisfaction for the assisted-living industry. Journal for Healthcare Quality, 23(2), 16-25.

Hall J, Milburn M, Roter D, Daltroy L. Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 1998;17(1):70–75.

Hudak, P. L. & J.G. Wright. (2000). The Characteristics of Patient Satisfaction Measures. Spine 25 (24): 3167-3177.

Institute of Medicine (2001). Improving the Quality of Long-Term Care, National Academy Press, Washington, D.C., 2001.

Kajonius, P. & Kazemi, A. (2016). Advancing the Big Five of user-oriented care and accounting for its variations. International Journal of Health Care Quality Assurance. 29(2): 162 – 176.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

Omnibus Budget Reconciliation Act (OBRA) of 1987. (1987, December 22). Public Law 100-203. Subtitle C: Nursing Home Reform.

Riccio, P.A. (2000). Quality Evaluaiton of home nursing care: Perceptions of patients, physicians, and nurses. Nursing Administration Quarterly 24(3): 43-52.

Shippee, T.P., Henning-Smith, C., Kane, R.L, & Lewis, T. (2015). Resident- and Facility-Level Predictors of Quality of Life in Long-Term Care. The Gerontologist. 55(4):643-655.

Uman, C & Urman, H. (1997). Measuring consumer satisfaction in nursing home residents. Nutrition 13: 705-707.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

The consumer movement has fostered the notion that patient evaluations should be an integral component of health care. Patient satisfaction, which is one form of patient evaluation, became an essential outcome of health care widely advocated for use by researchers and policy makers. Managed care organizations, accreditation and certification agencies, and advocates of quality improvement initiatives, among others, now promote the use of satisfaction surveys. For example, satisfaction information is included in the Health Plan Employer Data Information Set (HEDIS), which is used as a report card for managed care organizations (NCQA, 2016).

Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes. A 2014 systematic review and meta-analysis of randomized controlled trials, in which the patient-provider relationship was systematically manipulated and tracked with health care outcomes, found a small but statistically significant positive effect of the patient-provider relationship on health care outcomes (Kelly et al., 2014). This finding aligns with other studies that show a link between patient satisfaction and the following health-related behaviors:

1. Keeping follow-up appointments (Hall, Milburn, Roter, & Daltroy, 1998);
2. Disenrollment from health plans (Allen & Rogers, 1997); and,
3. Litigation against providers (Penchansky & Macnee, 1994).

The positive effect of person-centered care and patient satisfaction is not precluded from skilled nursing facilities. A 2013 systematic review of studies on the effect of person-centered initiatives in nursing facilities, such as the Eden Alternative, found person-centered care associated with psychosocial benefits to residents and staff, notwithstanding variations and limitations in study designs (Brownie & Nancarrow, 2013).

From the nursing facility and provider perspective, there are numerous ways to improve patient satisfaction. One study found conversations regarding end-of-life care options with family members improve overall satisfaction with care and increase use of advance directives (Reinhardt et al., 2014). Another found an association between improving symptom management of nursing home residents with dementia and higher satisfaction with care (Van Uden et al., 2013). Improvements in a nursing home food delivery system also were associated with higher overall satisfaction and improved resident health (Crogan et al., 2013). The advantage of the CoreQ: Long-Stay questionnaire is it is broad enough to capture patient dissatisfaction on various provided services and signal to providers to drill down and discover ways of improving the patient experience at their facility.

Specific to the CoreQ: Long-Stay questionnaire, the importance of the satisfaction areas assessed were examined with focus groups of residents and family members. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. Table 1c.5 in the appendix shows the score of the importance for questions included in the CoreQ: Long-Stay questionnaire. The overall ranking used was 10=Most important and 1=Least important. That the final three questions included in the measure had average scores ranging from 9.50 to 9.69 clearly shows that the respondents value the items used in the CoreQ: Long-Stay measure.

Allen HM, & Rogers WH. (1997). The Consumer Health Plan Value Survey: Round Two. Health Affairs. 1997;16(4):156–66.

Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. Clinical Interventions In Aging. 8:1-10.

Crogan, N.L., Dupler, A.E., Short, R., & Heaton, G. (2013). Food choice can improve nursing home resident meal service satisfaction and nutritional status. Journal of Gerontological Nursing. 39(5):38-45.

Hall J, Milburn M, Roter D, Daltroy L (1998). Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 17(1):70–75.

Kelley J.M., Kraft-Todd G, Schapira L, Kossowsky J, & Riess H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and metaanalysis of randomized controlled trials. PLoS One. 9(4): e94207.

Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care was generally high during 2005-09, with some variability across facilities. Health Affairs. 32(8):1416-25.

Lin, J., Hsiao, C.T., Glen, R., Pai, J.Y., & Zeng, S.H. (2014). Perceived service quality, perceived value, overall satisfaction and happiness of outlook for long-term care institution residents. Health Expectations. 17(3):311-20.

National Committee for Quality Assurance (NCQA) (2016). HEDIS Measures. http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures.aspx. Accessed March 2016.

Penchansky and Macnee, (1994). Initiation of medical malpractice suits: a conceptualization and test. Medical Care. 32(8): pp. 813–831

Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. Journal Of Social Work In End-Of-Life & Palliative Care. 10(2):112-26.

Van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. International Psychogeriatrics. 25(10):1697-707.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not Applicable

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator is the sum of the individuals in the facility that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long -Stay Resident questionnaire.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
A specific date is chosen. On that date all residents in the facility are identified. The data is then collected from all the residents in the facility meeting eligibility criteria on that date. Residents are given a maximum 2 month time window to complete the survey. While the frequency in which the questionnaires are administered is left up to the provider, they should at least administer the CoreQ questionnaire once a year. Last, only surveys returned within two months of the resident initially receiving the survey are accepted.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator includes all of the long-stay residents that had an average response =>3 on the CoreQ: Long Stay Resident questionnaire that do not meet any of the exclusions (see exclusions).

The calculation of an individual patient's average satisfaction score is done in the following manner:
-Respondents within the appropriate time window (see: S.5) and who do not meet the exclusions (See: S.11) are identified.
- A numeric score is associated with each response scale option on the CoreQ: Long-Stay Resident questionnaire (that is, Poor=1, Average=2, Good=3, Very Good=4, and Excellent=5).
- The following formula is utilized to calculate the individual's average satisfaction score. [Numeric Score Question 1 + Numeric Score Question 2 + Numeric Score Question 3]/3
-The number of respondents whose average satisfaction score >=3 are summed together and function as the numerator.

For residents with one missing data point (from the 3 items included in the questionnaire) imputation is used (representing the average value from the other two available questions). Residents with more than one missing data point, are not counted in the measure (i.e., no imputation is used for these residents since their responses are excluded). Imputation details are described in Section S.22.
No risk-adjustment is used (see S.13).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator includes all of the residents that have been in the SNF for 100 days or more regardless of payer status; who received the CoreQ: Long-Stay Resident questionnaire (e.g. people meeting exclusions do not receive the questionnaire), who responded to the questionnaire within the two month time window, who did not have the questionnaire completed by somebody other than the resident, and who did not have more than one item missing.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population includes all current individuals in the SNF on a given day who have been in the SNF for 100 days or more and respond to the CoreQ: Long-Stay Resident questionnaire and completed the survey within the two month time window (See: S.5).

Residents have up to 2 months to complete and return the survey. The length-of-stay is identified from nursing facility records (MDS item A1600 "Entry Date").

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclusions made at the time of sample selection are the following: (1) Residents who have poor cognition defined by the BIMS score; (2) residents receiving hospice; (3) residents with a legal court appointed guardian; and (4) residents who have lived in the SNF for less than 100 days.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (two months after the administration date) b) surveys that have more than one questionnaire item missing c) surveys

from residents who indicate that someone else answered the questions for the resident. (Note this does not include cases where the resident solely had help such as reading the questions or writing down their responses.)

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Individuals are excluded based on information from the Minimum Data Set (MDS) 3.0 assessment.

(1) Residents who have poor cognition: Then the Brief Interview for Mental Status (BIMS), a well validated dementia assessment tool is used. BIMS ranges are 0-7 (lowest); 8-12; and 13-15 (highest). Residents with BIMS scores of equal or less than 7 are excluded. (MDS Section C0200-C0500 items are used) (Saliba, et al., 2012).

(2) Patients receiving or having received any hospice. This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").

(3) Patients with court appointed legal guardian for all decisions will be identified from nursing facility health information system.

(4) Residents who have lived in the SNF for less than 100 days will be identified from the MDS. This is recorded in the MDS (Section A1600, Entry Date).

(5) Residents that respond after the 2 month response period (see S.18, section 3.a on how this is determined).

(6) Residents whose responses were completed by someone other than the resident will be excluded. Identified from an additional question on the CoreQ: Long-Stay Resident questionnaire.

(7) Residents without usable data (defined as missing data for 2 or 3 of the survey questions).

Saliba D, Buchanan J, Edelen MO, Streim J, Ouslander J, Berlowitz D, Chodosh J.
J Am Med Dir Assoc. 2012 Sep;13(7):611-7. doi: 10.1016/j.jamda.2012.06.004. Epub 2012 Jul 15.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
No stratification is used (see below).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not Applicable

S.16. **Type of score:**
Other (specify):
If other: Non-weighted score. Score is a percent.

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1.Identify the residents that have been residing in the SNF for 100 days or more. Length of stay so far is the MDS target date (TRGT_DT) - MDS admission date (A1900).

2. Take the residents that have been residing in the SNF for >=100 days and exclude the following:
a. Residents who have poor cognition defined as any residents with BIMS scores of 7 or lower. (MDS Section C0200-C0500 used) (Saliba, et al., 2012).
b. Patients receiving or having received any hospice. This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice"). c. Residents with Court appointed legal guardian for all decisions will be identified from nursing facility health information system.

3. Administer the CoreQ: Long-stay Resident questionnaire (See S.25) to these individuals. The questionnaire should be administered to all residents in the SNF after exclusions in step 2 above. Communicate that residents have four weeks to respond to the survey. Note, we will include surveys received up to two months from administration but specify four weeks to help increase response rate and completion within a timely manner. This also allows providers to use follow-up strategy at 4 weeks to get responses by the 8 week cut off.

4.Create a tracking sheet with the following columns:
i. Data Administered
ii. Data Response Received
iii. Time to Receive Response ([Date Response Received – Date Administered])

5.Exclude any surveys received after 2 months from administration.
6.Exclude responses not completed by the intended recipient (e.g. questions were answered by a friend or family members (Note: this does not include cases where the resident solely had help such as reading the questions or writing down their responses).

7.Exclude responses that are missing data for 2 or 3 of the CoreQ questions.

8.All of the remaining surveys are totaled and become the denominator.

9.Combine the CoreQ: Long-Stay Resident questionnaire items to calculate a resident level score. Responses for each item should be given the following scores:
a.Poor = 1,
b.Average = 2,
c.Good = 3,
d.Very Good =4 and
e.Excellent = 5.

10.Impute missing data if only one of the three questions are missing data.

11.Calculate resident score from usable surveys.
a.Patient score= (Score for Item 1 + Score for Item 2 + Score for Item 3) / 3.
i.For example, a resident rates their satisfaction on the three CoreQ questions as excellent = 5, very good = 4, and good = 3. The resident's total score will be 5 + 4 + 3 for a total of 12. The resident total score (12) will then be divided by the number of questions (3), which equals 4.0. Thus the residents average satisfaction rating is 4.0. Since the resident's score is >3.0, this resident will be counted in the numerator.
b.Flag those patients with a score equal to or greater than 3.0. These residents will be included in the numerator.

12. Calculate the CoreQ: Long-Stay Resident Measure which represents the percent of residents with average scores of 3.0 or above. CoreQ: Long-Stay Resident Measure= ([number of respondents with an average score of =3.0] / [total number of respondents])*100.

13.No risk-adjustment is used.

Saliba, D., Buchanan, J., Edelen, M.O., Streim, J., Ouslander, J., Berlowitz, D, & Chodosh J. (2012). MDS 3.0: brief interview for mental status. Journal of the American Medical Directors Association, 13(7): 611-617.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
No sampling is used. 100% residents not meeting exclusions are to receive the survey. No proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
1.Administer the CoreQ: Long-Stay Resident questionnaire to SNF residents who have resided in the SNF for >=100 days and who do not fall into one of the following exclusions:
a.Identify that the SNF resident has resided in the facility for >= 100 days. Using MDS (Section A1600, Entry Date).
b.Remove individuals with the following exclusions from the sample:
i.Residents who have poor cognition; Residents with BIMS scores of 7 are lower are excluded. (MDS Section C0200-C0500 used) (Saliba, et al., 2012).
ii.Patients receiving or having received any hospice. This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").
iii.Residents with Court appointed legal guardian for all decisions will be identified from nursing facility health information system.

2.Administer the CoreQ: Long-Stay Resident questionnaire to residents.

3.Instruct residents that they must respond to the survey within 2 months.

4.The response rate is calculated based on the number of usable surveys returned divided by the number of surveys administered.
a.As stated in S.11, surveys with missing responses for two or more questions, surveys received outside of the time window (more than two months after administration date), and surveys who were completed by someone else other than the intended resident are excluded
b.A minimum response rate of 30% needs to be achieved for results to be reported for a SNF.

5.Regardless of response rate, SNFs must also achieve a minimum number of 20 usable questionnaires (e.g. denominator). If after 2 months, less than 20 usable questionnaires are received then a facility level satisfaction measure is not reported.

6.All the questionnaires that are received (other than those with more than one missing value; or those returned after 2 months; or those completed by another person other than the intended resident) must be used in the calculations.

Saliba, D., Buchanan, J., Edelen, M.O., Streim, J., Ouslander, J., Berlowitz, D, & Chodosh J. (2012). MDS 3.0: brief interview for mental status. Journal of the American Medical Directors Association, 13(7): 611-617.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data was uncommon in the CoreQ: Long Stay Resident questionnaire testing (4.2% of any one of the 3 items). For residents with one missing data point (from the 3 items included in the questionnaire) imputation will be used (representing the

average value from the other available data points). As specified in S.11, residents to have more than one missing data point are excluded.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Healthcare Provider Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The collection instrument is the CoreQ: Long-Stay Resident questionnaire and exclusions are from the Resident Assessment Instrument Minimum Data Set (MDS) version 3.0.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
CoreQ_Long_Stay_Testing_Final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: CoreQ: Long-Stay Resident Measure
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***		☒ Outcome (*including PRO-PM*)
☐ Cost/resource		☐ Process
☐ Efficiency		☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If resident preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about resident preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on resident factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Resident preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of residents who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: CoreQ: Long-Stay Resident questionnaire	☒ other: CoreQ: Long-Stay Resident questionnaire, Pilot CoreQ: Long-Stay Resident questionnaire, Nursing Home Compare and CASPER

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

First, the Pilot CoreQ: Long-Stay Resident questionnaire containing an extended list of questions included on the CoreQ: Long-Stay Resident questionnaire was utilized for reliability and validity testing.

Second, data from the CoreQ: Long-Stay Resident questionnaire was used to test the measure for reliability and validity.

Third, to validate the measure, we also utilized Certification and Survey Provider Enhanced Reporting (CASPER) Quality Indicators and data form Nursing Home Compare.

1.3. What are the dates of the data used in testing? Click here to enter date range

June, 2014-September, 2014

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency

☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: Individual Resident

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The testing and analysis included three data sources, one of which had additional variables collected for a subset of respondents:

1. The Pilot CoreQ: Long-Stay Resident questionnaire was examined using responses from 1,714 residents from a national sample of nursing facilities.
    a. In addition, resident-level sociodemographic (SDS) variables were examined using this same sample of 1,714 residents (#1 above) in nursing facilities across the US.
2. Validity testing of the Pilot CoreQ: Long-Stay Resident questionnaire was examined using responses from 100 residents from the Pittsburgh area.
3. CoreQ: Long-Stay Resident measure was examined using 223 facilities and included responses from 7,307 residents. These nursing facilities were located in multiple states across the US.

Some basic descriptive characteristics of these facilities (data sources) are provided below in table 1.5.

Table 1.5: Descriptive Statistics of Centers Included in the Analyses

Data Source	Average Number of Licensed Beds	Average Daily Census	Sample Size of Residents (N)
Listed #1 (above)	139	121	1,714
Listed #2 (above)	202	188	100
Listed #3 (above)	137	130	7,307

1.6. How many and which residents were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of residents included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how residents were selected for inclusion in the sample*)

Resident Level of Analysis

Data was used from the CoreQ: Long-Stay Resident questionnaire. The questionnaire was administered to all eligible long-stay residents (with the exclusions described in the Specifications section of this application). The testing and analyses included:

1. The Pilot CoreQ: Long-Stay Resident questionnaire was examined using responses from 1,714 residents from a national sample of nursing facilities. (Data #1)
    a. In addition, resident-level sociodemographic (SDS) variables were examined using this same sample of 1,714 residents (Data #1 above) in nursing facilities across the US.
2. Validity testing of the Pilot CoreQ: Long-Stay Resident questionnaire was examined using responses from 100 residents from the Pittsburgh area. (Data #2)

3. CoreQ: Long-Stay Resident questionnaire MEASURE was examined using 223 facilities and included responses from 7,307 residents. These nursing facilities were located in multiple states across the US. (Data #3)

The descriptive characteristics of the residents are given in the following table that includes information from all of the data used (the education level and race information comes only from the sample described above with 1,714 respondents, as this data was not collected for the other samples).

Table 1.6: Patient Demographics (all samples pooled)

DEMOGRAPHICS		**Percent**
How long were you a resident at this facility?	<6 Months	12%
	6Months-1Yr	18%
	1-2Yrs	25%
	2-3Yrs	17%
	>3yrs	28%
Are you male or female?	Male	35%
	Female	65%
What year were you born?	Average	1931
What is the highest grade or level of school that you have completed?	Some HS	24%
	HS or GED	44%
	Some College/ 2yr Degree	20%
	4yr College Degree	7%
	>4yr College Degree	4%
What is your race?	White	86%
	Black	6%
	Asian	2%
	Native Hawaiian	0%
	American Indian	7%

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

We conducted two levels of testing in the development of the CoreQ: Long-Stay Resident measure. The first focused on testing (e.g., reliability, validity, exclusions) of the CoreQ: Long-Stay Resident questionnaire. The first source of data (pilot data) was utilized in developing and choosing the items to be included in the CoreQ: Long-Stay Resident questionnaire. This included using a questionnaire with 18 items. Below we call this the Pilot CoreQ: Long-Stay Resident questionnaire (i.e., Data #1, above). A subset of 100 residents from Data #1 was chosen in Data #2 to conduct a lagged re-administration of the same survey to measure agreement in response for the same resident regarding the same period of time.

Once the CoreQ: Long-Stay Resident questionnaire was developed, a second source of data was used to test the validity of the CoreQ: Long-Stay Resident measure (i.e., facility and summary score validity). This second data source is described above (i.e.223 facilities including responses from 7,307 residents [Data #3, above]).

1.8 What were the resident-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, resident-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each resident (e.g. census tract), or resident community characteristics (e.g. percent vacant housing, crime rate).

The following resident-level sociodemographic variables were available for analysis. For the distributions of these categories, see Tables 1.6 above.

- Age
  - Exact date of birth
- Sex
  - Male
  - Female
- Highest level of education
  - Some high school, but did not graduate
  - High school graduate or GED
  - Some college or 2 year degree
  - 4 year college graduate
  - More than 4 year college degree
- Race
  - White
  - Black or African American
  - Asian
  - Native Hawaiian or other Pacific Islander
  - American Indian or Alaskan Native.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

We measured reliability at the: (1) data element level; (2) the person/questionnaire level; and, (3) at the measure (i.e., facility) level. More detail of each analysis follows.

(1) DATA ELEMENT LEVEL

To determine if the CoreQ: Long-Stay Resident questionnaire items were repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, we re-administered the questionnaire to residents 1 month after their completion of the first survey. The Pilot CoreQ: Long-Stay Resident questionnaire had responses from 100 residents; we re-administered the survey to 50 of these same

residents. The re-administered sample was a sample of convenience as they represented residents from the Pittsburgh area (the location of the team testing the questionnaire). To measure the agreement, we calculated first the distribution of responses by question in the original round of surveys, and then again in the follow-up surveys (they should be distributed similarly); and second, calculated the correlations between the original and follow-up responses by question (they should be highly correlated).

(2) PERSON/QUESTIONNAIRE LEVEL

Having tested whether the data elements matched between the pilot responses and the re-administered responses, we then examined whether the person-level results matched between the Pilot CoreQ: Long-Stay Resident questionnaire responses and their corresponding re-administered responses. In particular, we calculated the percent of time that there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re-administered response was poor, average, good, very good or excellent.

(3) MEASURE (FACILITY) LEVEL

Last, we measured stability of the facility-level measure when the facility's score is calculated using multiple "draws" from the same population. This measures how stable the facility's score would be if the underlying residents are from the same population but are subject to the kind of natural sample variation that occurs over time. We did this by bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score calculated on the Pilot CoreQ: Long-Stay Resident questionnaire sample.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

(1) DATA ELEMENT LEVEL

Table 2a2.3.a shows the four CoreQ: Long-Stay Resident questionnaire items, and the response per item for both the pilot survey of 100 residents and the re-administered survey of 50 residents. The responses in the pilot survey are not statistically significant from the re-administered survey. This shows that the data elements were highly repeatable and produced the same results a high proportion of the time when assessing the same population in the same time period.

Table 2a2.3.a: CoreQ: Long-Stay Resident Questionnaire Responses from the Pilot and Re-administered Survey

Questionnaire Item	Response	Percent [Pilot Survey (N=100)]	Percent [Re-Administered Survey (N=50)]
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	4%	4%
	Average	12%	12%
	Good	30%	29%
	Very Good	28%	27%

	Excellent	20%	34%
2. Overall, how would you rate the staff?	Poor	2%	3%
	Average	11%	10%
	Good	31%	32%
	Very Good	31%	32%
	Excellent	21%	20%
3. How would you rate the care you receive?	Poor	2%	2%
	Average	12%	13%
	Good	32%	32%
	Very Good	28%	28%
	Excellent	21%	22%

NO SIGNIFICANT DIFFERENCES AT p=0.01

Table 2a2.3.b shows the average of the percent agreement from the first survey score to the second survey score for each item in the CoreQ: Long-Stay Resident questionnaire. This shows very high levels of agreement.

Table 2a2.3.b: Average Percent Agreement between the Pilot and Re-administered Survey

Questionnaire Item	Percent Agreement
4. In recommending this facility to your friends and family, how would you rate it overall?	97.6%
5. Overall, how would you rate the staff?	98.5%
6. How would you rate the care you receive?	98.0%

(2) PERSON/QUESTIONNAIRE LEVEL

Table 2a2.3.c shows the CoreQ: Long-Stay Resident questionnaire items, and the agreement in response per item for both the PILOT survey of 100 residents compared with the re-administered survey of 50 residents. The person-level responses in the PILOT survey are not statistically significant from the re-administered survey. This shows that a high percent of time there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re-administered response was poor, average, good, very good or excellent. Table 2a2.3.d shows the agreement between the pilot and re-administered responses. In summary, 97% or more of the re-administered responses agreed with their corresponding pilot responses, in terms of whether or not they were rated in the categories of poor or average or good, very good or excellent.

Table 2a2.3.c: Average Percent Agreement between Responses per Item for the Pilot Survey and Re-Administered Survey

Questionnaire Item	Response	Percent Person-Level Agreement in Response for the Pilot Survey (N=100) vs. Re-Administered Survey (N=50)
1. In recommending this facility to your friends	Poor	97%
	Average	97%
	Good	96%

and family, how would you rate it overall?	Very Good	98%
	Excellent	99%
2. Overall, how would you rate the staff?	Poor	98%
	Average	97%
	Good	98%
	Very Good	96%
	Excellent	99%
3. How would you rate the care you receive?	Poor	99%
	Average	99%
	Good	98%
	Very Good	97%
	Excellent	98%

Table 2a2.3.d: Average Percent Agreement between Response Options for the Pilot Survey and Re-Administered Survey

		Re-Administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
Pilot Response	Poor (1) or Average (2)	98.75%	98.5%
	Good (3), Very Good (4), or Excellent (5)	98.75%	99%

(3) MEASURE (FACILITY) LEVEL

After having performed the 10,000-repetition bootstrap, 14.18% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample, 20.91% were within 3 percentage points, 33.50% were within 5 percentage points, and 46.33% were within 10 percentage points.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

In summary, the measure displays a high degree of element-level, questionnaire-level, and measure (facility)-level reliability. First, the CoreQ: Long-Stay Resident questionnaire data elements were highly repeatable, with pilot and re-administered responses agreeing between 97% and 99% of the time depending on the question. That is, this produced the same results a high proportion of the time when assessed in the same population in the same time period. Second, the questionnaire level scores were also highly repeatable, with pilot and re-administered responses agreeing 98.5% of the time (or more). Third, a facility drawing residents from the same underlying population will only vary modestly. The 10,000-repetition bootstrap results show that the CoreQ: Long-Stay Resident measure scores from the same facility are moderately stable given the minimum sample size of 20 we set for this measure; and the maximum sample size was 122.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or

resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
In the development of the CoreQ: Long-Stay Resident questionnaire, three sources of data were used to perform three levels of validity testing. These are described above in Section 1.5.

The first source of data (data from a sample of convenience collected near the researchers developing the questionnaire in Pittsburgh) was used in developing and choosing the format to be utilized in the CoreQ: Long-Stay Resident questionnaire (i.e., response scale).

The second source of data, was pilot data collected from a national sample of 1,714 residents. This data was used in choosing the items to be used in the CoreQ: Long-Stay Resident questionnaire (i.e., questionnaire items). This data was also used in examining resident-level sociodemographic (SDS) variables.

The third source of data (collected from 223 facilities) was used examine the validity of the CoreQ: Long-Stay Resident measure (i.e., facility and summary score validity). These residents / nursing facilities were from multiple states across the U.S.

Thus, the following sections describe this validity testing:
1. Validity Testing of the questionnaire format used in the CoreQ: Long-Stay Resident questionnaire (using data source 1, from above);
2. Testing the items for the CoreQ: Long-Stay Resident questionnaire (using data source 2, from above);
3. Testing to determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Long-Stay Resident measure) (using data source 3, from above);
4. Validity testing for the CoreQ: Long-Stay Resident measure (also using data source 1, from above).

Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Resident Questionnaire

A. The face validity of the domains used in the CoreQ: Long-Stay Resident questionnaire was evaluated via a literature review. The literature review was conducted to examine important areas of satisfaction for LTC residents. Specifically, the research team examined 12 commonly used satisfaction surveys and reports to determine the most valued domains when looking at satisfaction. These surveys were identified by completing internet searches in PubMed and Google. Key terms that were searched included: resident satisfaction, long-term care satisfaction, and elderly satisfaction.

B. The face validity of the domains was also examined using a focus group of residents. The overall ranking used was 1=Most important and 22=Least important. That is residents were asked to rank the domains from most important to least important. The respondents were residents (N=40) in five nursing facilities in the Pittsburgh region.

C. The face validity of the Pilot CoreQ: Long-Stay Resident questionnaire response scale was also examined. The respondents were residents (N=40) in five nursing facilities in the Pittsburgh region. The percent of respondents that stated they "fully understood" how the response scale worked, could complete the scale, AND in cognitive testing understood the scale was used.

D. The Flesch-Kinkaid scale (Streiner & Norman, 1995) was used to determine if respondent correctly understood the questions being asked.

Streiner, D. L. & Norman, G.R. (1995). Health measurement scales: A practical guide to their development and use. 2nd ed. New York: Oxford.

1. Testing the Items for the CoreQ: Long-Stay Resident Questionnaire

The second series of validity testing was used to further identify items that should be included in the CoreQ: Long-Stay Resident questionnaire. This analysis was important, as all items in a satisfaction measure should have adequate psychometric properties (such as low basement or ceiling effects). For this testing, (1) A pilot group of 40 residents was first used in focus groups; (2) a Pilot version of the CoreQ: Long-Stay Resident questionnaire survey was administered consisting of 18 items (N= 1,714 residents). The testing consisted of:

A. Residents were asked to rate the 18 different satisfaction questions related to their experience in SNFs. This was conducted with a pilot group of 40 residents in focus groups.

B. The Pilot CoreQ: Long-Stay Resident questionnaire items performance with respect to the distribution of the response scale and with respect to missing responses. (using 1,714 residents described above)

C. The intent of the Pilot instrument was to have items that represented the most important areas of satisfaction (as identified above) in a parsimonious manner. Additional analyses such as exploratory factor analysis (EFA) were used to further refine the pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items. (using 1,714 residents described above)

3. To determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Resident Measure).

The CoreQ: Long-Stay Resident measure under development was meant to represent overall satisfaction with as few items as possible. The testing given below describes how this was achieved.

A. To support the construct validity that the idea that the CoreQ items measured a single concept of "satisfaction" – we performed a correlation analysis using all items in the instrument.

B. In addition, using all items in the instruments a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item was examined.

4. Validity Testing for the Core Q: Long-Stay Resident Measure.

A. To determine if the 3 items in the CoreQ: Long-Stay Resident questionnaire were a reliable indicator of satisfaction, the correlation between these three items (the "CoreQ: Long-Stay Resident Measure") and ALL of the items on the Pilot CoreQ instrument was conducted.

B. We performed additional validity testing of the facility-level CoreQ: Long-Stay Resident measure by examining the correlations between the CoreQ: Long-Stay Resident measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, and ii) several other quality metrics from Nursing Home Compare. If the CoreQ Long Stay Family scores correlate negatively with the measures that decrease as they get better, and positively with the measures that increase as they get better, then this supports the validity of the CoreQ Long Stay Family measure.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Resident Questionnaire

A. The face validity of the domains used in the CoreQ: Long-Stay Resident questionnaire was evaluated via a literature review (described above).

The research team examined the surveys and reports to identify the different domains that were included. The research team scored the domains by simply counting if an instrument included the domain. Table 2b2.3.a gives

the domains that were found throughout the search, as well as a score. An example is the domain clinical care, this was used in 10 out of the 12 surveys identified in the literature. An interpretation of this finding would be that items addressing clinical care are extremely important in satisfaction surveys. These domains were used in developing the pilot CoreQ: Long-Stay Resident questionnaire items.

Table 2b2.3.a: Survey Domain Score out of 12

Domain	Score out of 12	Domain	Score out of 12
Food	11	Spiritual	4
Activities	10	Confidence in Caregivers	3
Administration	10	Language and Communication	3
Clinical Care	10	Personal Suite	3
Staff Interaction	10	Therapy	3
Choice and Decision Making	9	Care Access	2
Facility Environment	9	Case Manager	2
Security and Safety	9	Comfort	2
Overall	8	Maintenance	2
Staff Overall	7	Move In	2
Autonomy and Privacy	6	Non-Clinical Staff Services	2
Housekeeping	6	Transitions	2
Personal Care	6	Transportation	2
Recommend facility	6	Emergency Response	1
Resident to Resident Friendships	5	Finances	1
Family Involvement	4	Time	1
Resident to Staff Friendships	4	Trust	1

B. The face validity of the domains was also examined using residents (described above). The following abbreviated table shows the rank of importance for each group of domains. The overall ranking used was 1=Most important and 22=Least important. The ranking of the 3 areas used in the CoreQ: Long-Stay Resident questionnaire are shown. Note, the food domain was ranked third – but was excluded from the CORE Q based on additional analyses showing that it was highly correlated with the overall domain; thus, it added little to the measure.

Table 2b2.3.b: Face Validity Abbreviated Results

Domain / Question	Average Rank
OVERALL (In recommending this facility to your friends and family, how would you rate it overall?)	2
STAFF (Overall, how would you rate the staff?)	1
CARE (How would you rate the care you receive?)	4

C. The face validity of the pilot CoreQ: Long-Stay Resident questionnaire response scale was also examined (described above). Table 2b2.3.c gives the percent of respondents that stated they "fully understood" how the response scale worked, could complete the scale, AND in cognitive testing understood the scale.

Table 2b2.3.c: Resident Understanding of Response Scale

Scale Format	Residents
Yes – No	100%
Yes – Somewhat – No	100%
Always – Usually – Sometimes –Never	100%
Very happy – Somewhat happy – Unhappy	100%
Excellent – Good – Fair – Poor	100%
Very Good – Good – Average – Poor – Very Poor	100%
Very Satisfied – Satisfied – Neither Satisfied or Dissatisfied – Dissatisfied – Very Dissatisfied	100%
4 Point Satisfaction Scale (1=Very unsatisfied, 2=Unsatisfied, 3=Neutral, 4=Satisfied)	100%
5 Point Likert Scale (1=Poor, 2=Average, 3=Good, 4=Very Good, 5=Excellent)	100%
Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree	95%
5 Point Importance Scale (1=Very important, 5=Very unimportant)	95%
5 Point Expectancy Scale (1=Not met, 2=Nearly met, 3=Met, 4=Exceeded, 5=Far exceeded expectations)	90%
10 Point Satisfaction Scale (1=Poor, 10=Excellent)	90%
8 Point Satisfaction Scale (1=Very dissatisfied, 2=Dissatisfied, 3=Somewhat dissatisfied, 4=Neither satisfied nor dissatisfied, 5=Somewhat satisfied, 6=Satisfied, 7=Very satisfied, 8=No response)	85%

Note: Highlighted cell represents the scale used in the CoreQ.

D. The CoreQ: Long-Stay Resident questionnaire was purposefully written using simple language. No *a priori* goal for reading level was set, however a Flesch-Kinkaid scale score of six, or lower, is achieved for all questions.

2. Testing the Items for the CoreQ: Long-Stay Resident Questionnaire

A. Each resident was asked to rate on a scale of 1 to 10 (with 10 as the best) how important they thought the question was for evaluating the experience with SNF care. The three questions included in the COREQ were highly rated out of all the questions and in analysis of resident's responses to 18 questions. That is, these three items were shown to provide unique information to distinguish satisfaction with SNFs. Specifically, "In recommending this facility to your friends and family, how would you rate it overall?" had an average score of 9.69; "Overall, how would you rate the staff?" had an average score of 9.56; and, "How would you rate the care you receive?" had an average score of 9.5. This shows a very pervasive influence of the satisfaction items with the experience of SNF care. See Table 1c.5 (Appendix).

B. The pilot CoreQ: Long-Stay Resident questionnaire items are shown in Table 2b2.3.d in the appendix. It also shows that the items performed well with respect to the distribution of the response scale and with respect to missing responses.

C. Using all items in the instruments (excluding the global item Q1 ("How would you rate the facility?")) exploratory factor analysis (EFA) was used to evaluate the construct validity of the measure. The Eigenvalues from the principal factors (unrotated) are presented in the Table below. In this analysis, the first Eigenvalue is overwhelmingly greater than the second Eigenvalue, this supports the proposition that the CoreQ instrument is measuring a single global concept of customer satisfaction – rather than a number of sub-concepts of customer satisfaction. Sensitivity analyses using principal factors and rotating provide highly similar findings.

Table 2b2.3.e: Exploratory Factor Analysis Results

	Long-Stay Resident
Factor 1	9.61
Factor 2	0.37

3. ***To determine if a Sub-Set of Items could be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Resident measure).***

A. To support the construct validity that the idea that the CoreQ items measured a single concept of "satisfaction" – we performed a correlation analysis using all items in the instrument. The analysis identifies the pairs of CoreQ items with the highest correlations. The highest correlations are shown in the Table 2b2.3.f. Items with the highest correlation are potentially providing similar satisfaction information. Note, the table provides 6 sets of correlations, the analysis was conducted examining all possible correlations between items. Because items with the highest correlation were potentially gathering similar satisfaction information they could be eliminated from the instrument.

Table 2b2.3.f: CoreQ: Long-Stay Resident Questionnaire Example Item Correlations

	Long-Stay
Highest Correlation	Q9-Q8 (.744)
Next highest Correlation	Q9-Q6 (.696)
Next highest Correlation	Q9-Q10 (.690)
Next highest Correlation	Q6-Q24 (.674)
Next highest Correlation	Q13-Q14 (.668)
Next highest Correlation	Q6-Q10 (.664)

RESULT = ITEMS TO DROP

C. In addition, using all items in the instrument a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item is shown in the table below. Cronbach's alpha measures the internal consistency of the values entered into the factor analysis; a value of 0.7 or higher is generally considered acceptably high. The additional item(s) is considered best in the sense that it is most highly correlated with the existing item, and therefore provides little additional information about the same construct. So this analysis was also used to eliminate items. Note, the table again provides 7 sets of correlations, the analysis was conducted examining all possible correlations between items. See table 2b2.3.g.

Table 2b2.3.g: Secondary Correlation Analysis of CoreQ: Long-Stay Resident Questionnaire Items

	Short-stay
Q1 + last satisfaction item ADD	Q6 (.854) Q10 (.852) Q9 (.847)
Q1 + ADD ADD	Q2 + Q6 (.853) Q9 + Q6 (.850) Q10 + Q6 (.847)
Q1 + ADD ADD	Q10 + Q9 (.858) Q10 + Q6 (.855) Q9 + Q6 (.854)

Thus, using the correlation information and factor analysis 3 items representing the CoreQ: Long-Stay Resident questionnaire were identified.

4. ***Validity Testing for the Core Q: Long-Stay Resident Measure.***

The overall intent of the analyses described above was to identify if a sub-set of items could reliably be used to produce an overall indicator of satisfaction, the CoreQ: Long-Stay Resident questionnaire.

A. The items were all scored according to the rules identified elsewhere. The same scoring was used in creating the 3 item CoreQ: Long-Stay Resident questionnaire summary score and the satisfaction score using the Pilot CoreQ: Long-Stay Resident questionnaire. The correlation was identified as having a value of 0.89. That is, the correlation score between actual the "CoreQ: Long-Stay Resident Measure" and all of the 18 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 3 items or the 18 item Pilot instrument.

B. We performed additional validity testing of the facility-level CoreQ: Long-Stay Resident measure by measuring the correlations between the CoreQ: Long-Stay Resident measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, and ii) several other quality metrics from Nursing Home Compare.

CoreQ: Long-Stay Resident measure is the percentage of residents discharged from the facility within 100 days of admission from a hospital to the nursing facility who, on average for the three CoreQ items included in the measure, rated the facility >= 3. We measured satisfaction using resident's responses to the three items from the CoreQ: Long-Stay Resident questionnaire (see Table 2a2.3.a).

The summary score from the 3 CoreQ: Long-Stay Resident questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the resident is calculated. The facility score represents the percent of residents with average scores of 3 or above. This score should be associated with quality. Therefore, for each facility in the sample the correlation with other quality indicators was examined.

(i) Relationship with CASPER Quality Indicators

Certification and Survey Provider Enhanced Reporting (CASPER) contains data collected as part of state/federal nursing home inspections. In short, nursing facilities that accept residents with Medicare and/or Medicaid payments are surveyed. This includes most (i.e., 97% [16,000 facilities]) nursing homes in the U.S. The survey process occurs approximately yearly, and includes the recording of many quality characteristics of the nursing home. These include restraint use; pressure ulcers; catheter use; antipsychotic use; antidepressant use; antianxiety use; and, use of hypnotics. These are commonly used quality indicators used for examining the quality of nursing homes.

In addition, when a nursing home is determined not to meet a certification minimum standard a deficiency citation is issued. These deficiency citations are also commonly used in the analyses of the quality of nursing homes. Approximately 180 deficiency citations exist and are grouped into 16 categories. These 16 categories group like areas together. They were developed by CMS and have considerable face validity; although, one limitation of using these categories is that they were not defined using empirical estimation (such as factor analysis).

Table 2b2.3.g: Correlation results between the CoreQ Long Stay Resident Questionnaire Measure Score and CASPER Quality Indicators

Quality Indicator	Correlation with Satisfaction Summary Score	P-Value
Any Deficiency Citations	-0.396	0.05
Physical Restraint Use	-0.105	0.12
Pressure ulcers	-0.105	0.12
Catheterized	-0.115	0.09
Antipsychotic medications	-0.152	0.02
Antidepressant medications	-0.472	0.05
Antianxiety medications	-0.149	0.03
Hypnotic medications	-0.476	0.05

(ii) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels

Nursing Home Compare (NHC) is a nursing home report card. After several years of pilot testing, the Centers for Medicare and Medicaid Services (CMS) released this report card on the world-wide web in November of 2002. Briefly, Nursing Home Compare provides information for facility location, structural factors (such as ownership), and staffing characteristics (such as registered nurse [RN] staffing levels). Most significantly, standardized quality information is presented in what are called Quality Measures (QMs). These are calculated from MDS information.

At the time period of for this study (i.e., 2014) CMS reported on 19 measures – these are called the core Quality Measures. The Quality Measures address specific areas of resident care, 5 are for short-stay residents and 14 are for long-stay residents. Long-stay measures are for those residents staying at a facility 3 months or more and short-stay measures are for residents staying at a facility less than 3 months. The long-stay measures are most pertinent to the CoreQ: Long-Stay Resident questionnaire; therefore, these were used in the analyses.

Nursing Home Compare also uses a five-star rating for facilities. This is based on information from the health inspection, direct care staffing, and the MDS quality measures. A five star facility is the highest score and a 1 star facility the lowest score. With respect to staffing, two measures are used: 1) RN hours per resident day; and 2) total staffing hours (RN+ LPN+ nurse aide hours) per resident day.

Table 2b2.3.h: Correlation Results between the CoreQ Long Stay Resident Questionnaire Measure Score and NHC Quality Indicators, Five Star Ratings, and Staffing Levels

Quality Indicator	Correlation with Satisfaction Summary Score MEASURE	P-Value
Percent of long-stay residents experiencing one or more falls with major injury.	-0.132	0.12

Percent of long-stay residents with a urinary tract infection	-0.209	0.08
Percent of long-stay residents who self-report moderate to severe pain	-0.206	0.05
Percent of long-stay high-risk residents with pressure ulcers	-0.320	0.05
Percent of long-stay low-risk residents who lose control of their bowels or bladder	-0.101	0.19
Percent of long-stay residents who have/had a catheter inserted and left in their bladder	-0.458	0.02
Percent of long-stay residents who were physically restrained	-0.211	0.04
Percent of long-stay residents whose need for help with daily activities has increased	-0.239	0.05
Percent of long-stay residents who lose too much weight	-0.122	0.10
Percent of long-stay residents who have depressive symptoms	-0.153	0.10
Percent of long-stay residents assessed and given, appropriately, the seasonal influenza vaccine	0.410	0.06
Percent of long-stay residents assessed and given, appropriately, the pneumococcal vaccine	0.333	0.05
Percent of long-stay residents who are administered antipsychotic medications	0.121	0.09
Five-Star rating	0.42	0.03
RN hours per resident day	0.47	0.05
Total staffing hours	0.39	0.04

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Resident Questionnaire

A. The literature review shows that domains used in the Pilot CoreQ: Long-Stay Resident questionnaire items have a high degree of both face validity and content validity.

B. Residents overall rankings, show the general "domain" areas used indicates a high degree of both face validity and content validity.

C. The results show that 100% of residents are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.

D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the item.

2. *Testing the Items for the CoreQ: Long-Stay Resident Questionnaire*

A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks etc.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument

as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

3. Testing to Determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Resident measure)

A. Using the correlation information of t*he Core Q: Long-Stay Resident questionnaire (18 items)* and the 3 items representing the CoreQ: Long-Stay Resident questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Long-Stay Resident Measure

A. The correlation of the 3 item CoreQ: Long-Stay Resident measure summary score (identified elsewhere in this document) with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.89.

That is, the correlation score between actual the "CoreQ: Long-Stay Resident Measure" and all of the 18 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 3 items or the 18 item Pilot questions.

This indicates that the CoreQ: Long-Stay Resident measure score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B.

(i) *Relationship with CASPER Quality Indicators*

The 8 CASPER Quality Indicators all had a reasonable level of negative correlation with the CoreQ: Long-Stay Resident measure in the direction as expected (higher satisfaction is associated with better quality. These correlations range from -0.105 to -0.476. The CoreQ: Long-Stay Resident measure is associated with these quality indicators. This testing indicates a reasonable degree of construct validity and convergent validity.

(i) *Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels*

The 13 Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels all had a moderate to high level of correlation and in the direction predicted with the CoreQ: Long-Stay Resident measure. These correlations range from ± 0.100 to 0.47. The CoreQ: Long-Stay Resident measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.47 for RN hours per resident-day and 0.37 for total staffing hours per resident day. This testing indicates a reasonable degree of construct validity and convergent validity.

As noted by Mor and associates (2003, p.41) "there is only a low level of correlation among the various measures of quality" In long term care settings. Castle and Ferguson (2010) also show the pattern of findings

of quality indicators in nursing facilities is consistently moderate with respect to the correlations identified. The magnitude of correlations of the CoreQ with quality metrics are consistent with these findings in this setting.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
To develop the CoreQ: Long-Stay Resident measure, we convened an expert panel to advise us on aspects such as which exclusions to apply to the measure with the goal to make sure as many residents who are capable of giving a response are included and that the voice of the resident is included not proxies.

The analysis of the impact exclusion had was performed on 223 nursing homes that have used the CoreQ: Long-Stay Resident measure. These facilities were included in multiple states across the US (this is data source 3, from above).

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
The expert panel advised us to exclude: 1) Residents with dementia impairing their ability to answer the questionnaire defined as having a low BIMS score; (2) residents receiving hospice care; and (3) Residents with a legal court appointed guardian.

[In addition we exclude; (4) Residents who have lived in the SNF for less than 100 days; (5) Respondents who have one or more missing data point (on the COREQ items); and (6) residents without usable data defined as missing data on 2 or 3 of the 3 questions.]

These exclusions are often used with satisfaction surveys (Sangl et al., 2007). Because the exclusions were based on individual's ability to answer questions and were also made in the pilot, we are not able to confirm if the exclusions actually made a difference to the scores, which is why we cannot calculate the mean CoreQ: Long-Stay Resident scores with and without the exclusions. However, the exclusions were made at the time of data collection, so we are able to report descriptive statistics regarding the number of exclusions made.

The exclusion analysis included responses from 223 facilities (described elsewhere). The exclusions were tracked and from these facilities included 34% of residents who have poor cognition; 2% residents with hospice; and 4% residents with a legal court appointed guardian.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. *Journal of Aging and Social Policy*, 19(2), 63-82.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If resident preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
These exclusions were applied because such residents were either unable to provide an independent response (e.g, residents who have poor cognition or a legal court appointed guardian) or for whom the burden of completing a questionnaire is inappropriate given their clinical situation and (e.g. hospice residents who are extremely sick and in the dying process), or residents whose answers we could not be confident were accurate or unbiased (residents who have poor cognition and durable power of attorney)). Therefore, the value of

excluding these residents takes into account burden on respondents and their ability to answer the questions. Thus, it is not possible to obtain answers or estimates of answers from non-respondents.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in resident characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

No research (to date) has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS®[1] Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical, theoretical, or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to resident characteristics and the satisfaction scores.

RTI International, Harvard University, RAND Corporation. *CAHPS Instrument for Persons Residing in Nursing Homes,* Final Report to CMS, CMS Contract No. CMS-01-01176, Sept. 2003.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select resident factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; resident factors should be present at the start of care*)

Not Applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not Applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Analyses used to examine SDS factors include: (1) the summary score for each of the 3 CoreQ: Long-Stay Resident questionnaire items; (2) the summary score for the CoreQ: Long-Stay Resident measure; and (3) the summary score from the CoreQ: Long-Stay Resident questionnaire measure (at the facility level).

(1) Summary Score for each of the 3 CoreQ: Long-Stay Resident Questionnaire Items

The summary score for each of the 3 CoreQ: Long-Stay Resident questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. Correlation and T-test a**nalyses were** used to compare the SDS means with each other (See 2b4.4b.a). These analyses show that the individual item scores used in the CORE Q: Long-Stay Resident measure are not significantly different based on either education level or race. That is, the educational related to the scores for individual items.

Table 2b4.4b.a: Mean CoreQ: Long-Stay Resident Distribution Item by Level of Education and Race

What is the highest grade or level of school that you have completed?	Respondents	Q1	Q2	Q3
		Mean	Mean	Mean
Some high school, but did not graduate	***24% (n=360)***	3.62	3.63	2.81
High school graduate or GED	***44% (n=647)***	3.63	3.71	2.86
Some college or 2 year degree	***20% (n=301)***	3.51	3.59	2.73
4 year college graduate	***7% (n=106)***	3.52	3.79	2.86
More than 4 year college degree	***4% (n=63)***	3.71	3.97	2.98
RANK CORRELATION		0.0201	0.0334	0.0066

RANK CORRELATION OF ITEMS WITH EDUCATION: NONE SIGNIFICANT AT p=0.05

Table 2b4.4b.a: Mean CoreQ: Long-Stay Resident Distribution Item by Level of Education and Race (continued)

What is your race?	Respondents	Q1	Q2	Q3
		Mean	Mean	Mean
White	***85% (n=1265)***	3.61	3.71	2.83
Black or African-American	**6% (n=86)**	3.30	3.33	2.69
Asian	**2% (n=24)**	3.71	3.67	2.86
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0
TWO-SAMPE T-TEST	1 vs. 2	2.67	3.43	1.16
	1 vs. 3	0.44	0.23	0.15
	2 vs. 3	1.17	1.49	0.75

RACE ITEMS: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(2) Summary Score for the CoreQ: Long-Stay Resident Measure

The summary score for each of the 3 CoreQ: Long-Stay Resident questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the resident is then calculated. Correlation and T-test a**nalyses were** used to compare the SDS means with each other (See Table 2b4.4b.b). These analyses show that the CORE Q: Long-Stay Resident measure score is not significantly different based on either education level or race of respondents. That is, the educational makeup of the respondents or the racial makeup of the respondents does not appear related to the measure score.

Table 2b4.4b.b: Mean CoreQ: Long-Stay Resident Distribution Measure by Level of Education and Race

What is the highest grade or level of school that you have completed?	Respondents	Measure Score
		Mean
Some high school, but did not graduate	***24% (n=360)***	3.84

High school graduate or GED	***44% (n=647)***	3.83
Some college or 2 year degree	***20% (n=301)***	3.79
4 year college graduate	***7% (n=106)***	3.80
More than 4 year college degree	***4% (n=63)***	3.87

RANK CORRELATION OF MEASURE SCORE WITH EDUCATION: NOT SIGNIFICANT AT p=0.05

Table 2b4.4b.b: Mean CoreQ: Long-Stay Resident Distribution Measure by Level of Education and Race (continued)

What is your race?		
	Respondents	Measure Score
		Mean
White	***85% (n=1265)***	3.84
Black or African-American	**6% (n=86)**	3.71
Asian	**2% (n=24)**	3.95
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0
American Indian or Alaskan Native	**0% (n=0)**	0
		p-value
TWO-SAMPLE T-TEST	1 vs. 2	0.12
	1 vs. 3	0.16
	2 vs. 3	0.75

RACE MEASURE SCORE: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(3) Summary score from the CoreQ: Long-Stay Resident Measure (at the facility level).

The summary score for each of the 3 CoreQ: Long-Stay Resident questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the resident is calculated. The facility score represents the percent of residents with average scores of 3 or above. A t-test a**nalysis was** used to compare the mean scores (See Table 2b4.4b.c). This analysis demonstrated the CORE Q: Long-Stay Resident measure is not significantly different based on either education level or race. That is, the educational makeup of the respondents or the racial makeup of the respondents does not appear related to the measure.

Table 2b4.4b.c: CoreQ: Long-Stay Resident Score with and without stratification for Education and Race

What is the highest grade or level of school that you have completed?	Respondents	Measure Score		
		Score with SDS Characteristic vs. Without Characteristic		
Some high school, but did not graduate	***24% (n=360)***	82.3	83.2	n.s

High school graduate or GED	***44% (n=647)***	83.5	83.5	n.s
Some college or 2 year degree	***20% (n=301)***	83.3	82.5	n.s
4 year college graduate	***7% (n=106)***	83.6	83.4	n.s
More than 4 year college degree	***4% (n=63)***	82.9	83.3	n.s

N.S. = Not significant at p=0.05

Table 2b4.4b.c: CoreQ: Long-Stay Resident Score with and without stratification by Education and Race (Continued)

What is your race?	Respondents	Measure Score (Mean)		
		Score with SDS Characteristic vs. Without Characteristic		
White	***85% (n=1265)***	83.5	83.2	n.s
Black or African-American	**6% (n=86)**	83.6	83.3	n.s
Asian	**2% (n=24)**	83.2	83.4	n.s
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0

N.S. = Not significant at p=0.05

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

Provide the statistical results from testing the approach to controlling for differences in resident characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable

2b4.9. Results of Risk Stratification Analysis:
Not Applicable

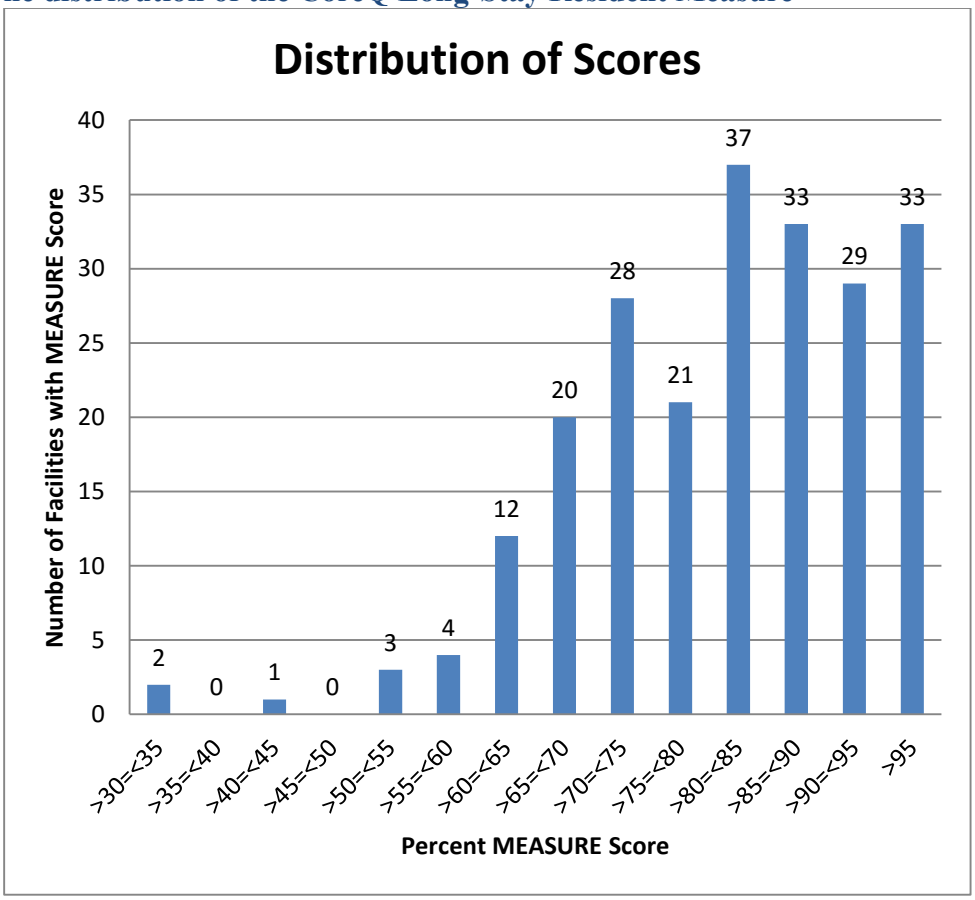
2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in resident characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted*)
Not Applicable
2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not Applicable

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b).*
We performed an analysis to examine whether the CoreQ Long-Stay Resident measure captured clinically/practically meaningful differences between providers by producing a histogram of the scores for the providers in the CoreQ: Long-Stay Resident questionnaire sample (Figure 2b5.2.1).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
The histogram below shows the distribution of the CoreQ Long-Stay Resident measure.

Figure 2b5.2.1: The distribution of the CoreQ Long-Stay Resident Measure



2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The CoreQ Long-Stay Resident scores reflect practical and meaningful differences in quality between facilities. First, the histogram in Section 2b5.2 shows that the distribution of summary scores is quite wide, indicating the scores can be used to differentiate facilities of varying levels of customer satisfaction quality.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not Applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS
2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Three items are used in the CoreQ: Long-Stay Resident questionnaire. In calculating the CoreQ: Long-Stay Resident measure if 1 item of 3 is missing then imputation is used, and if 2 (or more) of the 3 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. The testing to identify the extent and distribution of missing data included examining the frequency of missing responses for each of the 3 CoreQ: Long-Stay Resident questionnaire items and the extent and distribution of missing data for more than one missing response for the items. The method of testing to identify if the performance results were biased included examining the correlation with the quality indicators (described above) when imputation was and was not used.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of*

various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each)
As noted above, 3 items are used in the CoreQ: Long-Stay Resident questionnaire. In calculating the CoreQ: Long-Stay Resident measure if 1 item of 3 is missing then imputation is used, and if 2 (or more) of the 3 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. From the testing of 7,307 residents (described in section 1.5) we found:

1. In recommending this facility to your friends and family, how would you rate it overall?
That missing responses occurred in 4.86% (n=355) cases.
2. Overall, how would you rate the staff?
Missing responses occurred in 4.64% (n=339) cases.
3. How would you rate the care you receive?
Missing responses occurred in 4.56% (n=333) cases.

Two (or more) missing responses occurred in 123 cases. Thus, the degree of missing data was very small (=1.68%). Imputation was used in 904 cases or 12.37% of respondents.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Bias from imputation was minimal due to the rate of missingness being very low. The correlation with the quality indicators described above (i.e., restraint use, pressure ulcers, catheter use, antipsychotic use, antidepressant use, antianxiety use, use of hypnotics, and deficiency citations) was unchanged. When the respondents were removed from the analyses, the average Summary Scores remained the same.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Satisfaction Survey

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Since the CoreQ: Long-Stay Resident measure has been created and utilized in testing and quality improvement, we have modified it in the following ways.

We conducted analyses on collecting data for the suggested 2 month time period. Even the smallest nursing facilities were able to achieve the 20 survey response goal identified above. We identified that a majority of nursing facilities (i.e., 90%) in our sample could achieve this response rate if given 2 months. Therefore, this recommendation was incorporated into the specifications (given above).

As part of the CoreQ: Long-Stay Resident measure development, existing satisfaction vendors were contacted (including MyInnerView, Symbria, and NRC) for input on the administration and sample selection used. With respect to administration, the 2 month window used for including completed surveys are currently often used standard time periods used in the industry. With respect to the sample selection, the exclusion criteria (i.e., residents with court appointed legal guardian for all decisions; residents on hospice; residents who have poor cognition) were well received by these vendors. In many cases most of these sample selection criteria are already used by the vendors.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm) exist.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) AHCA Quality Initiative https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx Satisfaction Vendors N/A Quality Improvement (Internal to the specific organization) Large Nursing Home Chain N/A

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The CoreQ: Long-Stay Resident measure is currently in use by a large nursing home chain for the purposes of quality improvement. The data described above was collected from 223 facilities in this chain and included responses from 7,307 residents. These nursing facilities are located in multiple states across the US.

In addition, 10 large national satisfaction vendors in the SNF area have agreed to add the CoreQ to their questionnaires and calculate the measure. The following Customer Satisfaction Vendor are using CoreQ:
•Align
•Brighton Consulting Group
•Healthcare Academy (ReadyQ)
•inQ Experience Surveys
•National Research Corporation (My Innerview)
•Pinnacle
•Providigm/abaqis
•Sensight Surveys
•Service Trac
•The Jackson Group, Inc.

We do not have counts of patients being surveyed and geographical representation from the vendors, however they represent the majority of customer satisfaction vendors currently doing SNF business in the United States.

A letter has been sent to all 10,000 AHCA SNF members indicating which vendors to date have agreed to add the CoreQ to their questionnaire and calculate the measure (see attached letter in appendix, section 4.a.1). A user's manual has been developed and is available on AHCA's website for all satisfaction survey vendors to use. One of the vendors has added the CoreQ to their questionnaire used by states for mandatory satisfaction data collection in all their SNFs (RI, KS and GA), though the results have not yet been calculated by these states.

AHCA and NCAL have also incorporated the CoreQ into their national Quality Initiative goals. AHCA represents nearly 10,000 of the 15,000 SNFs and provides feedback to all of its members on their satisfaction scores using the CoreQ. This has resulted in growing number of members and vendors collecting the data.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The CoreQ: Long-Stay Resident measure is not currently publicly reported or used in other accountability applications (e.g., payment program, certification, licensing). The reason for this is that it is a new measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
AHCA has recently started the second Quality Initiative, laying out a series of quality improvement and reporting goals for the AHCA membership, which covers nearly 10,000 of all 15,000+ Medicare & Medicaid certified SNFs in the U.S. Among these goals is the collection and reporting of CoreQ customer satisfaction data. Because it has been included in the Quality Initiative 2015-2018, AHCA's machinery for publicizing and encouraging the adoption of the tool has been activated, including AHCA's quality division spending a large number of staff hours working to accomplish this. In addition to marketing the use of the survey instrument as a way for SNFs to understand how their patients view the care and other services that they were provided by the SNFs, AHCA is developing an upload and reporting feature within its member data profiling tool, LTC Trend TrackerSM, which allows SNFs to centrally view a large number of quality, compliance, operational and financial metrics from public and non-public sources. The CoreQ report and upload feature within LTC Trend Tracker will include an API for vendors performing the survey on behalf of SNFs – AHCA's preferred approach to collecting the data – so that the aggregate CoreQ results will be immediately available to providers as they are collected. Given that LTC Trend TrackerSM is the leading method for SNFs to profile their quality and other data, the incorporation of CoreQ into LTC Trend Tracker means it will immediately become the de facto standard for customer satisfaction surveys for the SNF industry.

In addition, large national satisfaction vendors in the SNF area, have agreed to add the CoreQ to their questionnaires and calculate the measure. An email has been sent to all 10,000 AHCA SNF members indicating which vendors to date have agreed to add the CoreQ to their questionnaire and calculate the measure (see attached letter in Section 4a.1 of the Appendix).

We also are working with states who require satisfaction measurement to incorporate the CoreQ into their process. The State of RI pilot tested a version of the CoreQ in its statewide satisfaction questionnaire for Long-Stay residents. The state of Massachusetts has included the CoreQ short stay as part of its current ongoing deliberation on measuring satisfaction in SNFs. AHCA has a presence in each state, and our state affiliates will be promoting the use of the CoreQ in those states that are collecting or considering collecting satisfaction.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not Applicable.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no negative consequences to individuals or populations identified during testing or evidence of unintended negative consequences to individuals or populations reported since the implementation of the CoreQ: Long-Stay Resident questionnaire or the measure that is calculated using this questionnaire. This is consistent with satisfaction surveys in general in nursing facilities. Many other satisfaction surveys are used in nursing facilities with no reported unintended consequences to patients or their families.
There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make them further dissatisfied.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0692 : Consumer Assessment of Health Providers and Systems (CAHPS®) Nursing Home Survey: Long-Stay Resident Instrument

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
The CoreQ: Long-Stay Resident measure does not conceptually address the same measure focus as any other NQF-endorsed measures, however it does conceptually address the same target population as another NQF-endorsed measure.

The Consumer Assessment of Health Providers and Systems (CAHPS®) Nursing Home Survey: Long-Stay Resident Instrument (NQF #0692) presented by the Agency for Healthcare Research and Quality received NQF approval over 4 years ago in Jan 24, 2012. This instrument is endorsed to collect resident satisfaction information and consists of a 50 item questionnaire. Our application also uses nursing home residents (The CoreQ: Long-Stay Resident measure) but consists of three items. No analyses have been conducted with CAHPS® such that a score representing satisfaction can be calculated. Whereas the CoreQ items are used to calculate this satisfaction score. Thus, the score from these items is used to provide standardized information on the overall resident satisfaction of the facility. The current CAHPS survey is not used in this way.

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The CoreQ: Long-Stay Resident measure does not conceptually address the same measure focus as any other NQF-endorsed measures, however it does conceptually address the same target population as another NQF-endorsed measure. The Consumer Assessment of Health Providers and Systems (CAHPS®) Nursing Home Survey: Long-Stay Resident Instrument (NQF #0692) presented by the Agency for Healthcare Research and Quality received NQF approval over 4 years ago in Jan 24, 2012. This instrument is endorsed to collect resident satisfaction information and consists of a 50 item questionnaire. Our application also uses nursing home residents (The CoreQ: Long-Stay Resident measure) but consists of three items. No analyses have been conducted with CAHPS® such that a score representing satisfaction can be calculated. Whereas the CoreQ items are used to calculate this satisfaction score. Thus, the score from these items is used to provide standardized information on the overall resident satisfaction of the facility. The current CAHPS survey is not used in this way.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** CoreQ_Long_Stay_Appendix_Final-635950196480014539.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Health Care Association
Co.2 **Point of Contact:** Urvi, Patel, upatel@ahca.org, 202-898-2858-
Co.3 **Measure Developer if different from Measure Steward:** American Health Care Association
Co.4 **Point of Contact:** Lindsay, Schwartz, lshwartz@ncal.org, 202-898-2848-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The workgroup gave input, reviewing our suggested administration, required response rate, the manual, and exclusions.

Mary Tess Crotty, Genesis - Also helped provide feedback on the development process and the user manual. Additionally, she reviewed the analyses.
Matt O'Connor HCR Manor Care- Also helped provide feedback on the development process and the user manual. Additionally, he conducted some analyses and reviewed the analyses.
Judy Hoff, Health Care Academy
Rich Kortum, My Innerview/National Research Corporation
Peter Kramer, abaqis/Providigm

Ellen Kuebrich, abaqis/Providigm
Michael Johnson, ServiceTrac
Chris Magelby, Pinnacle

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2015
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.6 **Copyright statement:** None
Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2616
Measure Title: CoreQ: Long-Stay Family Measure
Measure Steward: American Health Care Association
Brief Description of Measure: The measure calculates the percentage of family or designated responsible party for long stay residents (i.e., residents living in the facility for 100 days or more), who are satisfied (see: S.5 for details of the timeframe). This consumer reported outcome measure is based on the CoreQ: Long-Stay Family questionnaire that has three items.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.
The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.

As part of the Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay Family questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities

(Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long Stay Family measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay Family measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long-Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

Numerator Statement: The numerator assesses the number of family or designated responsible party for long stay residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party members for long stay residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long-Stay Family questionnaire.

Denominator Statement: The target population is family or designated responsible party members of a resident residing in a SNF for at least 100 days. The denominator includes all of the individuals in the target population who respond to the CoreQ: Long-Stay Family questionnaire within the two month time window (see S.5) who do not meet the exclusion criteria (see S.10).
Denominator Exclusions: Please note, the resident representative for each current resident is initially eligible regardless of their being a family member or not. Only one primary contact per resident should be selected.

Exclusions made at the time of sample selection include: (1) family or designated responsible party for residents with hospice; (2) family or designated responsible party for residents with a legal court appointed guardian; (3) representatives of residents who have lived in the SNF for less than 100 days; and (4) representatives who reside in another country.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (more than two months after the administration date) and b) surveys that have more than one questionnaire item missing.

Measure Type: PRO
Data Source: Healthcare Provider Survey
Level of Analysis: Facility

New - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This is a patient-reported outcome measure of family satisfaction. The developer provides a diagram and a table demonstrating the links between structures and/or processes and the outcomes that have been found to influence family satisfaction, and the final patient reported outcome of satisfaction.
- The developer notes that "Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management"
- The developer states "We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care" and notes that measuring patient satisfaction is required for person-centered care for three reasons:
    - Measuring satisfaction is necessary to understand patient preferences.
    - Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
    - Satisfaction information can help facilities improve the quality of care they provide

Guidance from the Evidence Algorithm

PRO-based measure (Box 1) → Relationship between the outcome and at least one healthcare action is identified and supported by the rationale (Box 2) → PASS

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provided the following information on performance gap:

- Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes.
- Studies show a link between patient satisfaction and the following health-related behaviors:
  - Keeping follow-up appointments
  - Disenrollment from health plans
  - Litigation against providers
- Family members are influential participants in the care of long stay patients in nursing home and thus gauging their satisfaction is also important.

The developer provided performance scores based on 6,192 family member responses from 221 facilities. A table of performance scores for facilities that met the inclusion criteria (20 valid responses and 50% response rate) is included. The facility score represents the percent of residents with average scores of 3 or above.

Facility Level Performance Distribution

Survey Item	Obs	Mean	Std. Dev.	Min	Max
coreq2	150	3.775434	.3636211	2.625	4.678571
coreq3	150	3.693261	.3712719	2.4375	4.714286
coreq1	150	3.618075	.3898564	2.3125	4.586207

Overall Descriptive Information for the Summary Score MEASURE

	min	p25	p50	p75	max
Summary Score	27.1	37.5	82.9	88.9	100

Disparities

The developer says differences in scores based on SDS categories were not statistically significant:

- By race/ethnicity, whites averaged a score of 83.47, Blacks or African-Americans averaged a score of 83.3, and Asians 83.5.
- By highest education level those with those high school but who did not graduate averaged 83.4, high school graduates averaged 83.3, those with some college or a 2-year degree averaged 82.5, 4 year college graduates averaged 83.2, and those with more than 4 year college degree averaged 83.6.
- By age group, residents younger than 65 years old averaged 71.7, those 65-74 averaged 83.7, those 75-84 averaged 87.6, and those older than 85 averaged 74.9.
- By gender, males averaged a score of 80.1 and females averaged a score of 86.1.

However, research over the last 20 years has consistently found poorer care in facilities with high minority populations and that nursing homes remain segregated, with black patients concentrated in poorer-quality homes (as measured by staffing ratios, performance, and are more financially vulnerable).

The measure is not risk adjusted.

Meaningfulness to the Target Population (PRO-PM):

- The developer provided an overview: "Specific to the CoreQ: Long Stay Family questionnaire, the importance of the satisfaction areas assessed were examined with focus groups of residents and family members. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. Table 1c.5 in the appendix shows the score of the importance for question included in the CoreQ: Long Stay Family questionnaire. The

overall ranking used was 10=Most important and 1=Least important. The final three questions included in the measure had average scores ranging from 9.5 to 9.69; this clearly shows that the respondents value the items used in the CoreQ: Long Stay Family measure."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence Supporting the Measure

Comments:

**There has been a call for more experience measures and there is correlation with overall satisfaction with the drivers presented.
**Yes. The PRO information is actionable to improve SNF care.
**Measures a PRO - there is a noted relationship between the PRO and at least one healthcare action. (There are steps/changes the providers can make based on results, i.e. patient-provider relationship)
**This PRO links directly to desired outcome of understanding patient preferences, helping consumers choose and trust a health care facility and accelerate facilities ability to improve

1b. Performance Gap

Comments:

**There was variation in results showing opportunity.
**Yes. Prior performance measures have been withdrawn, leaving a need for SNF evaluation.
**Yes, the measure demonstrates quality issues/concerns which will provide quantitative data for staff to make improvements.
There is also a need for national benchmarking which could be met by using this measure.
**Yes- patient satisfaction is related to triple aim- specifically, follow up apts, enrollment, med-legal litigations.

Disparities: measure is not risk adjusted. Research shows disparities in nursing homes with lager populations of low income and minority. Nursing homes also remain segregated with Blacks in lower quality homes (based on staffing, performance and finances)

1c. PROM-PRO

Comments:

**There is a mention that a focus group of patients and family members was done but it only mentions the number of patients in the focus group (40) and doesn't comment on the participation of family.
**100 family members/representatives completed the measure, followed by 50 re-surveyed one month later. Focus groups were also conducted to determine the meaningfulness to respondents. (This was on a small group of 40).

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): CoreQ: Long-Stay Family questionnaire; for exclusions the Resident Assessment Instrument Minimum Data Set (MDS) version 3.0 is used

Specifications:

- The level of analysis is facility.
- The measure result is a non-weighted percentage score:
    - The numerator assesses the number of family or designated responsible party for long stay residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party members for long stay residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long-Stay Family questionnaire.
    - The denominator includes all of the individuals in the target population (family or designated responsible party members of a resident who does not meet the exclusion criteria and who is residing in a SNF for at least 100 days) who respond to the CoreQ: Long-Stay Family questionnaire within the two month time window
    - There is no data dictionary.
- A calculation algorithm is described.
- The measure is not risk adjusted or stratified.
- There are 4 exclusions from the sample, and two added after survey administration:
    - family or designated responsible party for residents with hospice;
    - family or designated responsible party for residents with a legal court appointed guardian;
    - representatives of residents who have lived in the SNF for less than 100 days;
    - representatives who reside in another country.
    - surveys received outside of the time window (more than two months after the administration date) (excluded after administration)
    - surveys that have more than one questionnaire item missing (excluded after administration)
- The calculation of exclusion criteria is specified and includes MDS and nursing home facility health information system data.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data elements were tested using a test-retest methodology. The Pilot CoreQ Long Stay Family survey was administered to 100 family members/representatives; 50 were re-surveyed one month later. The distribution of responses and the correlation between the original and follow-up scores were then calculated.
- Person/questionnaire level was tested using the same test-retest methodology.
- The stability of the facility-level score was tested using bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score.

Results of reliability testing

Results for each level of testing are presented.

- **Data element testing** showed very high levels of agreement and no statistically significant difference in the responses to each question between the original and re-test results.

Average Percent Agreement between the Pilot and Re-administered Surveys

Questionnaire Item	Percent Agreement
1. In recommending this facility to your friends and family, how would you rate it overall?	97.1%
2. Overall, how would you rate the staff?	98.8%
3. How would you rate the care your family member received?	97.5%

- **Person/questionnaire level** agreement showed very high levels of agreement and no statistically significant difference in the responses to each question

Average Percent Agreement between Response Options for the Pilot Survey and Re-Administered Survey

		Re-Administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
Pilot Response	Poor (1) or Average (2)	98.5%	98.8%
	Good (3), Very Good (4), or Excellent (5)	98.5%	98.7%

- **Measure level testing** also demonstrated moderate agreement**:**
  - 11.5% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample
  - 20.9% were within 3 percentage points
  - 30.4% were within 5 percentage points
  - 42.2% were within 10 percentage points

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empiric testing- yes (box 2) -> with measure score – yes (box 4) – appropriate method – yes (box 5) – Level of certainty or confidence in the performance measure scores (box 6): Moderate

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *In review of the bootstrap analysis (measure level) across all three CoreQ measures, the results above show less agreement – is this a cause for concern about the reliability of the survey collected from family/relations?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

1. Validity testing of the questionnaire format used in the CoreQ: Long Stay Family questionnaire
   - Face validity evaluated via literature review and review of 12 commonly used satisfaction surveys; also examined face validity of domains and the response scale, using 40 patients in 5 nursing homes. The Flesch-Kinkaid scale was used to determine if patients understood the questions.

2. Testing the items for the CoreQ: Long Stay Family questionnaire;
   - Exploratory factor analysis (EFA) were used to further refine the pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items.

3. To determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Long Stay Family measure);
   - Correlation analysis and a factor analysis conducted on items

4. Validity Testing for the CoreQ: Long Stay Family measure.
   - Developers examined correlation between the three items in the measure and all of the items on the pilot instrument.
   - Also examined correlations between the CoreQ: Long Stay Family measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, ii) several other quality metrics from Nursing Home Compare

Validity testing results:

Results for each level of validity testing are provided. The developer interpretation of results is as follows:

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long Stay Family Questionnaire

   A. The literature review shows that domains used in the Pilot CoreQ: Long-Stay Family questionnaire items have a high degree of both face validity and content validity.

B. Family's overall rankings, show the general "domain" areas used indicates a high degree of both face validity and content validity.

C. The results show that 100% of Family's are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.

D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the item.

2. Testing the Items for the CoreQ: Long Stay Family Questionnaire

A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks etc.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

3. Determine if a Sub-Set of Items Could Reliably be Used to Produce an Overall Indicator of Satisfaction (The Core Q: Long Stay Family Measure).

A. Using the correlation information of the Core Q: Long-Stay Family questionnaire (18 items) and the 3 items representing the CoreQ: Long-Stay Family questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Long-Stay Family Measure

A. The correlation of the 3 item CoreQ: Long-Stay Family measure summary score (identified elsewhere in this document) with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.90.

That is, the correlation score between actual the "CoreQ: Long-Stay Family Measure" and all of the 18 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 3 items or the 18 item Pilot questions.

This indicates that the CoreQ: Long-Stay Family measure score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B.

(i) Relationship with CASPER Quality Indicators

The CASPER Quality Indicators all had negative correlation with the CoreQ: Long-Stay Family measure as expected (higher satisfaction is associated with better quality). These correlations range from ± 0.03 to 0.28. The CoreQ: Long-Stay Family measure is associated with these quality indicators. This testing indicates a reasonable degree of construct validity and convergent validity.

(ii) Relationship with Nursing Home Compare (*NHC) Quality Indicators, Five Star ratings, and staffing levels*

The Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels had a moderate to high level of correlation with the CoreQ: Long-Stay Family measure. These correlations range from ± 0.11 to 0.45. The CoreQ: Long-Stay Family measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.45 for RN hours per resident-day and 0.42 for total staffing hours per resident day. This testing indicates a reasonable degree of construct validity and convergent validity.

(iii) Relationship with the risk-adjusted Discharge to Community Measure

The risk-adjusted Discharge to community measure was negatively correlated to the CoreQ: Long-Stay Family measure. The correlations range from -0.03 to -0.06, all of which are not statistically significant at the p-value of 0.05. This was not as hypothesized which may be related to some SNFs that specialize in long stay, have very low discharge to community rates as admissions do not have a plan to go home.

(iv) Relationship with the risk adjusted PointRight® Pro 30™ Rehospitalizations

The risk-adjusted PointRight® Pro 30™ Rehospitalizations was negatively correlated to the CoreQ: Long-Stay

Family measure. The correlations range from -0.18 to -0.21, and all of them were statistically significant at the p-value of 0.05. This is expected because lower rehospitalization rates (an indicator of high quality) are associated with higher satisfaction scores. This was as hypothesized. This testing indicates a reasonable degree of construct validity and convergent validity.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- An expert panel advised the developer on exclusions. 1) Family members of residents receiving hospice care; and (2) Family members of residents with a legal court appointed guardian. In addition the developer excludes; (3) Family members of residents who have lived in the SNF for less than 100 days; (4) Respondents who have one or more missing data point (on the COREQ items); and (5) surveys received outside of the time window (more than two months after the administration date); all three are commonly excluded on satisfaction surveys.
- The first analysis included data from 221 facilities. Exclusions were tracked and the following reported:
  - 2% Family members of residents with hospice;
  - 4% family members with a legal court appointed guardian.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Risk adjustment summary

The developers provide the following rationale for no risk-adjustment:
"No research to date has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS®[1] Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical, theoretical or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to family characteristics and the satisfaction scores.
RTI International, Harvard University, RAND Corporation. *CAHPS Instrument for Persons Residing in Nursing Homes,* Final Report to CMS, CMS Contract No. CMS-01-01176, Sept. 2003."

Questions for the Committee:

- *A justification for no risk adjustment is provided. Is there any evidence that contradicts the developer's rationale and analysis?*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer states:

- We performed an analyses to examine whether the CoreQ Long-Stay Family measure captured clinically/practically meaningful differences between providers by examining a histogram of the scores for the providers in the CoreQ: Long-Stay Family questionnaire sample.
- Of the 221 facilities in the test population, scores ranged from 1 facility scoring 30-35% to 32 facilities scoring greater than 95%.

Question for the Committee:

Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer states missing data was uncommon (4.25-4.31% each for the three questions, and 3.8% for two or more missing responses). For patients with one missing data point (from the 3 items included in the CoreQ: Long Stay Family questionnaire) imputation is utilized (representing the average value from the other available data points); imputation was used in 3.5% of cases. Patients with more than one missing data point are excluded.

Preliminary rating for validity: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Guidance from the Validity Algorithm

Specifications consistent with evidence (Box 1): Yes →Potential threats to validity assessed (Box 2): Yes →Empirical validity testing performed using measure as specified (Box 3): Yes →Validity testing with computed performance measure score (Box 6): Yes →Method Described appropriate (box 7): Yes →Level of certainty or confidence that the performance measure score is a valid indicator of quality (Box 8): High

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 Specifications

Comments:

**This measure could be consistently implemented.

**No inconsistency noted.

**Does long stay include bounce back from ED/IP to SNF? Or, is it 100 days straight? Is palliative care also excluded or just hospice? Does family include non-wedded partners? (LGBTQ community?)

Specifications are consistent with evidence of what target populations finds meaningful

2a.2 Validity Testing

Comments:

**Patient level and facility both validated with follow-up / re-admistered surveys with consistent results.

**The test sample seemed low. /Reliability in the sample seemed OK. / Reliability of survey data from family members seemed reliable enough.

**Scope seems to small . Also, what was the family relationships within the focus group? The test re-test method was employed to 100 family members/reps and only 50 were re-tested. How were these 50 chosen? Random? The results of the reliability testing were very high but I wonder of the N is to small and how the re-test population was chosen. I am also unclear of case mix. How does disease acuity impact family satisfaction if at all? What about levels of caregiver stress?

2b.2 Validity Testing

Comments:
**The domains used have high correlation with longer surveys and with overall satisfaction.
**The test sample seems potentially low for generalization to the larger population.
Conclusions may be drawn from the results.
The score is a measure of quality, in that the PRO information provides input regarding performance at the SNF.
**Some question about the measure level testing which demonstrated 'moderate agreements' ???
**Testing method: face validity and empirical validity testing of the measure score

I agree that results demonstrate sufficient validity. Correlation between the coreQ and all of the 18 items used in the pilot instrument so that the satisfaction information is appx the same which indicates the COREQ ling stay measure does adequately represent the overall satisfaction of the facility

Flesh-Kinkaid scale score implemented to ensure high degree of understanding among family

low percent of missing responses

wide distribution

2b.3-2b.7 Validity Testing
Comments:
**No adjustments for SES made - reason given was lack of research that the results vary based on these factors.
**I wondered why Patients on Hospice care were excluded? Is this because they receive different care from other SNF patients?
Why are results excluded that have more than one missing answer. Were the non-answer results compared to those that completed all or all but one of the questionnaires. Could the non-answer be an answer?
**NA - agree with preliminary rating for validity of HIGH.
**I would say caregiver stress, social and economic barriers are risks associated with satisfaction and family characteristics
examined a histogram of the scores for the providers in the sample- normal distribution

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The collection instrument is the CoreQ: Long Stay Family questionnaire and Resident Assessment Instrument Minimum Data Set (MDS) version 3.0.
- This is a patient satisfaction survey conducted via mailed survey.
- No fees required to use the measure; the developer did not indicate if there are fees associated with the use of the survey.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility
Comments:

**67% of the facilities tested had a response rate of over 50%. Of note, for the resident surveys, the response rate for being included was over 30%.

**Not sure the data elements apply in this PRO.
This survey is not available electronically and will be mailed to respondents.
The data collection strategy can easily be operationalized.

**This measure seems feasible; no concerns. / Unsure why the preliminary rating is moderate?

**There is routine measurement of patient satisfaction currently used in care delivery

None of these are available within the electronic health record

Concerns:
surveys sent via mail are hard to get back
does this exclude population with poor literacy
language?/translation?
who reviews the data?
Does this exclude people who have extended stays but over various visits? For example the population who goes between SNF to acute care and generate a lot of IP and SNF days?

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

- AHCA Quality Initiative: https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx
- Satisfaction Vendors (10 national companies)

Quality Improvement (Internal to the specific organization)

- Large Nursing Home Chain

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details
Not in use for accountability program, but ACHA plans to begin public reporting of the CoreQ measures as part of their Quality Initiative 2016-2018 (9,600 SNFs)

Improvement results
N/A

Unexpected findings (positive or negative) during implementation
None reported

Potential harms

The developer states, “There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make them further dissatisfied.”

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**Intended for public reporting in the future.
**The performance results should help SNF's determine areas in which they need to improve.
**Not currently publicly reported; will be in the next few years (Quality Initiative 2016-2018).
**NO plan to use in accountability platform-
will not be publically reported

ACHA plans to begin public reporting of the CoreQ as part of the quality initiative

I DO think this should be publically reported as a way to empower families and caregivers to be active participants in reporting on the quality of skilled care- especially as this population grows rapidly. This will also help for continuous improvement of facilities (If they results are put to process improvement efforts). I also see benefits on recruitment and staff morale. Staff retainment is very hard in skilled facilities and being able to build a culture of customer service and excellent and drive to this will metrics is important

No potential harms cited. I can see harm in the definition of "family". Can this extend beyond the traditional definition of family now that there is more tolerance and acceptance for non-traditional families and caregivers. For example, for the LGBTQ community, this could be unintended harm.

Criterion 5: Related and Competing Measures

Related or competing measures

The developers cited potential relatedness/competing with a measure based on the CAHPS Nursing Home surveys, however; the measures derived from Nursing Home CAHPS have recently lost endorsement. AHRQ has communicated lack of resources to maintain the measures, and they are not currently in use in any federal program.

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: *CoreQ: Long-Stay Family Measure*
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: **Instructions**

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome
☒Patient-reported outcome (PRO): Customer Satisfaction

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Family satisfaction can be looked at as the outcome for a number of structures and processes within skilled nursing care centers. Drivers for high satisfaction rates include competency of staff, care/concern of staff, and responsiveness of management (National Research Corporation, 2014).

received.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the Health Professions,* 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

National Research Corporation. (2014). 2014 National Research Report Empowering Customer-Centric Healthcare Across the Continuum.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

The table below provides the structure and process drivers that influence long stay family satisfaction.

Authors	***Structure or Process and Driver of Family Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Reinhardt, et al., 2014	Process Care/concern of staff and competency of staff	Conversations regarding end-of-life care options with family members show higher overall satisfaction with care and more use of advance directives.	Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. *Journal Of Social Work In End-Of-Life & Palliative Care*. 10(2):112-26.

Lin et al., 2014.	Process Competency of staff	Significant difference for overall resident satisfaction with higher perceived service quality.	Lin, J., Hsiao, C.T., Glen, R., Pai, J.Y., & Zeng, S.H. (2014). Perceived service quality, perceived value, overall satisfaction and happiness of outlook for long-term care institution residents. *Health Expectations*. 17(3):311-20.
Van Uden et al. (2013).	Process Responsiveness of management	For nursing home residents with dementia improved symptom management is associated with higher satisfaction with care.	van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. *International Psychogeriatrics.* 25(10):1697-707.
Li et al. (2013).	Structure Responsiveness of management	Higher overall nursing home satisfaction scores were associated with higher nursing staffing levels and fewer deficiency citations.	Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care was generally high during 2005-09, with some variability across facilities. *Health Affairs*. 32(8):1416-25.
Authors	***Structure or Process and Driver of Family Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Crogan et al. (2013).	Process Responsiveness of management	Improvements in a nursing home food delivery system were associated with higher overall satisfaction and improved resident health.	Crogan, N.L., Dupler, A.E., Short, R., & Heaton, G. (2013). Food choice can improve nursing home resident meal service satisfaction and nutritional status. *Journal of Gerontological Nursing*. 39(5):38-45.
Brownie & Nancarrow (2013).	Structure & Process Responsiveness of management and care/concern of staff	Implementation of person-centered care is associated with higher levels of satisfaction.	Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. *Clinical Interventions In Aging*. 8:1-10.
Kleijer et al.,	Process	Residents perceive a low level of quality	Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts,

2014	Competency of staff	of care in centers where there is a high level of antipsychotic use.	A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.
Bishop et al., 2008	Structure Care/concern of staff	CNA's that receive a good supervision are more committed to staying in their jobs. This commitment in turn leads to positive relationships with resident and higher resident satisfaction.	Bishop, C., Weinberg, D., Leutz, W., Dossa, A., Pfefferle, S., & Zincavage, R. (2008). Nursing assistants' job commitment: Effect of nursing home organizational factors and impact on resident well-being. *The Gerontologist*, 48(1), 36-45.
Authors	***Structure or Process and Driver of Family Satisfaction***	***Summary Statement showing structures, processes, interventions and services and influence short-stay discharge satisfaction.***	***Citation***
Kayser-Jones et al., 1999	Structure Responsiveness of management and care/concern of staff	Higher levels of RN and LPN staffing have been associated with better quality outcomes such as ADL maintenance and hydration. Centers that have a family council in addition to the required resident council have higher resident satisfaction.	Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision. *Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. *Journal of Aging and Social Policy*, 19(2), 9-42.

Donabedian, A. (1985). Twenty years of research on the quality of medical care: 1964-1984. *Evaluation and the Health Professions,* 8, 243-65.

Donabedian, A. (1988). The quality of care. *Journal of the American Medical Association,* 260, 1743-1748.

Donabedian, A. (1996). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44(1), 166-203.

Glass, A. (1991). Nursing home quality: A framework for analysis. *Journal of Applied Gerontology*, 10(1), 5-18.

Kleijer, B., Van Marum, R., Frijeters, D., Jansen, P., Ribbe, M., Egberts, A., & Heerdink, E. (2014). Variability between nursing homes in prevalence of antipsychotic use in patients with dementia. *International Psychogeriatrics,* 26(3), 363-371.

Bishop, C., Weinberg, D., Leutz, W., Dossa, A., Pfefferle, S., & Zincavage, R. (2008). Nursing assistants' job commitment: Effect of nursing home organizational factors and impact on resident well-being. *The Gerontologist*, 48(1), 36-45.

Lucas, J.A., Lowe, T.J., Robertson, B., Akincigil, A., Sambamoorthi, Q., Bilder, S., Paek, E.K., & Crystal, S. (2007). The relationship between organizational factors and resident satisfaction with nursing home care and life. *Journal of Aging & Social Policy,* 19(2), 125-151.

Kayser-Jones, J., Schell, E.S., Poter, C., Barbaccia, J.C., & Shaw, H. (1999). Factors contributing to dehydration in nursing homes: Inadequate staffing and lack of professional supervision. *Journal of the American Geriatrics Society*, 47(10), 1187-1194.

Kane, R.L., & Kane, R.A. (2001). What older people want from long-term care, and how can they get it. *Health Affairs*, 20(6), 114-127.

Westat. Resident experience with nursing home care: A literature review.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2616
De.2. **Measure Title:** CoreQ: Long-Stay Family Measure
Co.1.1. **Measure Steward:** American Health Care Association
De.3. **Brief Description of Measure:** The measure calculates the percentage of family or designated responsible party for long stay residents (i.e., residents living in the facility for 100 days or more), who are satisfied (see: S.5 for details of the timeframe). This consumer reported outcome measure is based on the CoreQ: Long-Stay Family questionnaire that has three items.
1b.1. **Developer Rationale:** Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.
The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.

As part of the Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay Family questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long

Stay Family measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay Family measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long-Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

S.4. **Numerator Statement:** The numerator assesses the number of family or designated responsible party for long stay residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party members for long stay residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long-Stay Family questionnaire.
S.7. **Denominator Statement:** The target population is family or designated responsible party members of a resident residing in a SNF for at least 100 days. The denominator includes all of the individuals in the target population who respond to the CoreQ: Long-Stay Family questionnaire within the two month time window (see S.5) who do not meet the exclusion criteria (see S.10).
S.10. **Denominator Exclusions:** Please note, the resident representative for each current resident is initially eligible regardless of their being a family member or not. Only one primary contact per resident should be selected.

Exclusions made at the time of sample selection include: (1) family or designated responsible party for residents with hospice; (2) family or designated responsible party for residents with a legal court appointed guardian; (3) representatives of residents who have lived in the SNF for less than 100 days; and (4) representatives who reside in another country.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (more than two months after the administration date) and b) surveys that have more than one questionnaire item missing.

De.1. **Measure Type:** PRO
S.23. **Data Source:** Healthcare Provider Survey
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not Applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
CoreQ_Family_Evidence_Final-635950343462644989.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Collecting satisfaction information from skilled nursing facility (SNF) patients is more important now than ever. We have seen a philosophical change in healthcare that now includes the patient and their preferences as an integral part of the system of care. The Institute of Medicine (IOM) endorses this change by putting the patient as central to the care system (IOM, 2001). For this philosophical change to person-centered care to succeed, we have to be able to measure patient satisfaction for these three reasons:

(1) Measuring satisfaction is necessary to understand patient preferences.
(2) Measuring and reporting satisfaction with care helps patients and their families choose and trust a health care facility.
(3) Satisfaction information can help facilities improve the quality of care they provide.

The implementation of person-centered care in SNFs has already begun, but there is still room for improvement. The Centers for Medicare and Medicaid Services (CMS) demonstrated interest in consumers' perspective on quality of care by supporting the development of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey for patients in nursing facilities (Sangl et al., 2007).
Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Advancing Excellence in America's Nursing Homes campaign (2006), which lists person-centered care as one of its goals; Action Pact, Inc., which provides workshops and consultations with nursing facilities on how to be more person-centered through their physical environment and organizational structure; and Eden Alternative, which uses education, consultation, and outreach to further person-centered care in nursing facilities. All of these initiatives have identified the measurement of resident satisfaction as an essential part in making, evaluating, and sustaining effective clinical and organizational changes that ultimately result in a person-centered philosophy of care.
The importance of measuring resident satisfaction as part of quality improvement cannot be stressed enough. Quality improvement initiatives, such as total quality management (TQM) and continuous quality improvement (CQI), emphasize meeting or exceeding "customer" expectations. William Deming, one of the first proponents of quality improvement, noted that "one of the five hallmarks of a quality organization is knowing your customer's needs and expectations and working to meet or exceed them" (Deming, 1986). Measuring resident satisfaction can help organizations identify deficiencies that other quality metrics may struggle to identify, such as communication between a patient and the provider.

As part of the Department of Commerce renowned Baldrige Criteria for organizational excellence, applicants are assessed on their

ability to describe the links between their mission, key customers, and strategic position. Applicants are also required to show evidence of successful improvements resulting from their performance improvement system. An essential component of this process is the measurement of customer, or resident, satisfaction (Shook & Chenoweth, 2012).

The CoreQ: Long Stay Family questionnaire can strategically help nursing facilities achieve organizational excellence and provide high quality care by being a tool that targets a unique and growing patient population. Moreover, improving the care for long stay nursing home patients is tenable. A review of the literature on satisfaction surveys in nursing facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many nursing facilities by improving care (i.e., changing either structural or process aspects of care). This was based on satisfaction scores ranging from 60 to 80% on average.

It is worth noting, few other generalizations could be made because existing instruments used to collect satisfaction information are not standardized. Thus, benchmarking scores and comparison scores (i.e., best in class) were difficult to establish. The CoreQ: Long Stay Family measure has considerable relevance in establishing benchmarking scores and comparison scores.

This measure's relevance is furthered by recent federal legislative actions. The Affordable Care Act of 2010 requires the Secretary of Health and Human Services (HHS) to implement a Quality Assurance & Performance Improvement Program (QAPI) within nursing facilities. This means all nursing facilities have increased accountability for continuous quality improvement efforts. In CMS's "QAPI at a Glance" document there are references to customer-satisfaction surveys and organizations utilizing them to identify opportunities for improvement. Lastly, the new "Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities" proposed rule includes language purporting the importance of satisfaction and measuring satisfaction. CMS states "CMS is committed to strengthening and modernizing the nation's health care system to provide access to high quality care and improved health at lower cost. This includes improving the patient experience of care, both quality and satisfaction, improving the health of populations, and reducing the per capita cost of health care." There are also other references in the proposed rule speaking to improving resident satisfaction and increasing person-centered care (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015). The CoreQ: Long Stay Family measure has considerable applicability to both of these initiatives.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS (2009). Skilled Nursing Facilities Non Swing Bed - Medicare National Summary. http://www.cms.hhs.gov/MedicareFeeforSvcPartsAB/Downloads/NationalSum2007.pdf.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Deming, W.E. (1986). Out of the crisis. Cambridge, MA. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Institute of Medicine (2001). Improving the Quality of Long-Term Care. National Academy Press, Washington, D.C., 2001.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

MedPAC. (2015). Report to the Congress: Medicare Payment Policy. http://www.medpac.gov/documents/reports/mar2015_entirereport_revised.pdf?sfvrsn=0.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home residents. Journal of Aging and Social Policy, 19(2), 63-82.

Shook, J., & Chenoweth, J. (2012, October). 100 Top Hospitals CEO Insights: Adoption Rates of Select Baldrige Award Practices and Processes. Truven Health Analytics. http://www.nist.gov/baldrige/upload/100-Top-Hosp-CEO-Insights-RB-final.pdf.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data*

source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
The data source included 221 nursing facilities from multiple states across the US. The data were collected from June 2014 through Sept 2014, leading to responses from 6,192 family members or designated responsible party. The performance measure scores are available in section 1b.2 in the appendix, section 1b.2. This shows, on the 0 – 100 scale used for the CoreQ: Long-Stay Family measure (expressed in percent), the minimum score is 27.1, the 25th percentile is 37.5, the 50th percentile is 82.9, the 75th percentile is 88.9, and the maximum score is 100.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not Applicable.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We did not risk adjust the measure by sociodemographic status due to no statistically significant differences (at the 5% level) in the scores between the SDS categories. See Table 2b4.4b.b in the Testing section. By race, whites averaged a score of 83.47, Blacks or African-Americans averaged 83.3, and Asians 83.5; there were no observations for Native Hawaiians or other Pacific Islanders, American Indian or Alaskan Natives (Table 2b4.4b.c in the Testing section). By highest level of education, those with some high school but who did not graduate averaged 83.4, high school graduates averaged 83.3, those with some college or a 2 year degree averaged 82.5, 4 year college graduates averaged 83.2, and those with more than 4 year college degree averaged 83.6 (Table 2b4.4b.c in the Testing section). By age group, those younger than 65 years old averaged 71.7, those 65-74 averaged 83.7, those 75-84 averaged 87.3, and those older than 85 averaged 74.9 (Table 1b.4.a in the Appendix). Furthermore, by gender, males averaged a score of 80.1 and females averaged a score of 86.1 (Table 1b.4.a in the Appendix).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Multiple studies in the past twenty years have examined racial disparities in the care of nursing facility residents and have consistently found poorer care in facilities with high minority populations (Fennell et al., 2000; Mor et al., 2004; Smith et al., 2007). Work on disparities in quality of care between elderly white and black residents within nursing facility has shown clearly that nursing homes remain relatively segregated, and that nursing home care can be described as a tiered system in which blacks are concentrated in marginal-quality homes (Li, Ye, Glance & Temkin-Greener, 2014; Fennell, Feng, Clark & Mor, 2010; Li, Yin, Cai, Temkin-Greener, Mukamel, 2011; Chisholm, Weech-Maldonado, Laberge, Lin, & Hyer, 2013; Mor et al., 2004; Smith et al., 2007). Such homes tend to have serious deficiencies in staffing ratios, performance, and are more financially vulnerable (Smith et al, 2007; Chisholm et al., 2013). Based on a review of the nursing facility disparities literature, Konetzka and Werner (2009) concluded that disparities in care are likely related to racial and socioeconomic segregation as opposed to within-provider discrimination. This conclusion is supported, for example, by Grunier and colleagues who found that as the proportion of black residents in the nursing home increased the risk of hospitalization among all residents, regardless of race, also increased (Grunier et al., 2008). Thus, adjusting for racial status, has the unintended effect of adjusting for poor quality providers not to differences due to racial status.

We hypothesize that the blacks who tend to receive care in poor facilities would have lower satisfaction scores related to the overall quality in the SNF rather than differences in care blacks received compared to other ethnicities in the SNF, indicating that the best measure of racial disparities in satisfaction rates is one that measures scores at the facility level. That is, ethnic and social economic status differences are related to inter-facility differences not to intra-facility differences in care. Therefore, we believe the literature suggests that racial status should not be risk adjusted otherwise, one is adjusting for the poor quality of the SNFs rather than differences due to racial status.

In addition, even with the concentration of certain ethnicities in SNFs, the sample size for African Americans divided across all the nursing facilities also would make most nursing facilities unable to report a rate stratified by race (see below for state sample size).

Grabowski, D.C. (2004). The admission of Blacks to high-deficiency nursing homes. Medical Care 42(5): 456-464.

Gruneir, A., Miller, S. C., Feng, Z., Intrator, O., & Mor, V. (2008). Relationship between state Medicaid policies, nursing home racial composition, and the risk of hospitalization for black and white residents. Health Services Research, 43(3), 869-881.

Konetzka, R. T., & Werner, R. M. (2009). Review: Disparities in long-term care building equity into market-based reforms. Medical Care Research and Review, 66(5), 491-521.

Mor, V., Zinn, J., Angelelli, J., Teno, J. M., & Miller, S. C. (2004). Driven to tiers: socioeconomic and racial disparities in the quality of nursing home care. Milbank Quarterly, 82(2), 227-256.

Smith, D. B., Feng, Z., Fennell, M. L., Zinn, J. S., & Mor, V. (2007). Separate and unequal: racial segregation and disparities in quality across US nursing homes. Health Affairs, 26(5): 1448-1458.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

The definition of quality in a nursing facility has shifted from a focus on structure and process criteria to clinical outcomes, resident satisfaction, and quality of life. This shift was first supported by nursing home reform legislation included in the Omnibus Budget Reconciliation Act of 1987 (OBRA, 1987). Furthering the movement, the Institute of Medicine (IOM) put the patient as central to the care system (Castle, 2007; IOM, 2001) – necessitating the collection of satisfaction information. As mentioned previously (see 1b.1), a focus on person-centered care and satisfaction is also evident in the Quality Assurance & Performance Improvement Program (QAPI) for nursing facilities and proposed Reform Requirements for Long-Term Care Facilities (Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities, 2015).

Measuring and reporting satisfaction of nursing home care is important in many ways. First, residents are more likely to follow medical advice when they rate their care as satisfactory (Hall, Milburn, Roter, & Daltroy, 1998). Second, because resident satisfaction can influence the quality of care provided and the outcomes of treatment (Hudak and Wright 2000), satisfaction surveys can be used as measures of clinical and organizational accountability. Third, measuring and reporting resident satisfaction can help nursing facilities identify and improve aspects of quality. Furthermore, if publicly released, information on satisfaction with care can help elders and their families choose a nursing facility.

Several research efforts have concluded consumer satisfaction is an important indicator of quality of care in nursing homes (Gesell, 2001; Bangerter et al. 2016; Shippee et al 2015; Kajonius and Kazemi, 2016). In addition, other studies have concluded nursing resident satisfaction data provides information about quality of care that is different from clinician perspectives and clinical indicators (Berlowitz, Du, Kazis, & Lewis, 1993; Riccio 2000; Uman & Urman, 1997). This exemplifies the need for resident satisfaction data to achieve person-centered care. Only by hearing from the patient can we ensure the care provided is person-centered.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Bangerter, L.R., Heid, A.R., Abbott, K, & Van Haitsma, K. (2016). Honoring the Everyday Preferences of Nursing Home Residents: Perceived Choice and Satisfaction with Care. The Gerontologist. (Advance online publication): 1-8.

Berlowitz, D. R., Du, W., Kazis, L., & Lewis, S. (1995). Health-related quality of life of nursing home residents: Difference in patient and provider perceptions. Journal of the American Geriatric Society, 43, 799-802.

Castle, N.G. (2007). A literature review of satisfaction instruments used in long-term care settings. Journal of Aging and Social Policy, 19(2), 9-42.

CMS, University of Minnesota, and Stratis Health. QAPI at a Glance: A step by step guide to implementing quality assurance and performance improvement (QAPI) in your nursing home. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/QAPI/Downloads/QAPIAtaGlance.pdf.

Gesell, S.B. (2001). A measure of satisfaction for the assisted-living industry. Journal for Healthcare Quality, 23(2), 16-25.

Hall J, Milburn M, Roter D, Daltroy L. Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 1998;17(1):70–75.

Hudak, P. L. & J.G. Wright. (2000). The Characteristics of Patient Satisfaction Measures. Spine 25 (24): 3167-3177.

Institute of Medicine (2001). Improving the Quality of Long-Term Care, National Academy Press, Washington, D.C., 2001.

Kajonius, P. & Kazemi, A. (2016). Advancing the Big Five of user-oriented care and accounting for its variations. International Journal of Health Care Quality Assurance. 29(2): 162 – 176.

Medicare and Medicaid Programs; Reform of Requirements for Long-Term Care Facilities; Department of Health and Human Services. 80 Fed. Reg. 136 (July 16, 2015) (to be codified at 42 CFR Parts 405, 431, 447, et al.).

Omnibus Budget Reconciliation Act (OBRA) of 1987. (1987, December 22). Public Law 100-203. Subtitle C: Nursing Home Reform.

Riccio, P.A. (2000). Quality Evaluaiton of home nursing care: Perceptions of patients, physicians, and nurses. Nursing Administration Quarterly 24(3): 43-52.

Shippee, T.P., Henning-Smith, C., Kane, R.L, & Lewis, T. (2015). Resident- and Facility-Level Predictors of Quality of Life in Long-Term Care. The Gerontologist. 55(4):643-655.

Uman, C & Urman, H. (1997). Measuring consumer satisfaction in nursing home residents. Nutrition 13: 705-707.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

The consumer movement has fostered the notion that patient evaluations should be an integral component of health care. Patient satisfaction, which is one form of patient evaluation, became an essential outcome of health care widely advocated for use by researchers and policy makers. Managed care organizations, accreditation and certification agencies, and advocates of quality improvement initiatives, among others, now promote the use of satisfaction surveys. For example, satisfaction information is included in the Health Plan Employer Data Information Set (HEDIS), which is used as a report card for managed care organizations (NCQA, 2016).

Measuring and improving patient satisfaction is valuable to patients, because it is a way forward on improving the patient-provider relationship, which influences health care outcomes. A 2014 systematic review and meta-analysis of randomized controlled trials, in which the patient-provider relationship was systematically manipulated and tracked with health care outcomes, found a small but statistically significant positive effect of the patient-provider relationship on health care outcomes (Kelly et al., 2014). This finding aligns with other studies that show a link between patient satisfaction and the following health-related behaviors:

1. Keeping follow-up appointments (Hall, Milburn, Roter, & Daltroy, 1998);
2. Disenrollment from health plans (Allen & Rogers, 1997); and,
3. Litigation against providers (Penchansky & Macnee, 1994).

The positive effect of person-centered care and patient satisfaction is not precluded from skilled nursing facilities. A 2013 systematic review of studies on the effect of person-centered initiatives in nursing facilities, such as the Eden Alternative, found person-centered care associated with psychosocial benefits to residents and staff, notwithstanding variations and limitations in study designs (Brownie & Nancarrow, 2013).

Moreover, family members are influential participants in the care of long stay patients in nursing home and thus gauging their satisfaction is also important. For instance, a study found that “relatives [of nursing home patients] attributed responsibility for most tasks to nursing home staff but held themselves responsible for monitoring and evaluating quality of care, teaching staff to deliver high quality care, and providing direct care intended to preserve the residents ´self´” (Bowers, 1988). This is resonated by the CoreQ: Long Stay Family questionnaire items which assess overall satisfaction, satisfaction with the staff and the care that the family member received.

From the nursing facility and provider perspective, there are numerous ways to improve patient and family satisfaction. One study found conversations regarding end-of-life care options with family members improve overall satisfaction with care and increase use of advance directives (Reinhardt et al., 2014). Another found an association between improving symptom management of nursing home residents with dementia and higher satisfaction with care (Van Uden et al., 2013). Improvements in a nursing home food delivery system also were associated with higher overall satisfaction and improved resident health (Crogan et al., 2013). The advantage of the CoreQ: Long Stay Family questionnaire is it is broad enough to capture patient dissatisfaction on various provided services and signal to providers to drill down and discover ways of improving the patient experience at their facility.

Specific to the CoreQ: Long Stay Family questionnaire, the importance of the satisfaction areas assessed were examined with focus groups of residents and family members. The respondents were patients (N=40) in five nursing facilities in the Pittsburgh region. Table 1c.5 in the appendix shows the score of the importance for question included in the CoreQ: Long Stay Family questionnaire. The overall ranking used was 10=Most important and 1=Least important. The final three questions included in the measure had average scores ranging from 9.5 to 9.69; this clearly shows that the respondents value the items used in the CoreQ: Long Stay Family measure.

Allen HM, & Rogers WH. (1997). The Consumer Health Plan Value Survey: Round Two. Health Affairs. 1997;16(4):156–66.

Brownie, S. & Nancarrow, S. (2013). Effects of person-centered care on residents and staff in aged-care facilities: a systematic review. Clinical Interventions In Aging. 8:1-10.

Bowers, B. (1988). Family Perceptions of Care in a Nursing Home. The Gerontologist. 28(3): 361-368.

Crogan, N.L., Dupler, A.E., Short, R., & Heaton, G. (2013). Food choice can improve nursing home resident meal service satisfaction and nutritional status. Journal of Gerontological Nursing. 39(5):38-45.

Hall J, Milburn M, Roter D, Daltroy L (1998). Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol. 17(1):70–75.

Kelley J.M., Kraft-Todd G, Schapira L, Kossowsky J, & Riess H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and metaanalysis of randomized controlled trials. PLoS One. 9(4): e94207.

Li, Y., Cai, X., Ye, Z., Glance, L.G., Harrington, C., & Mukamel, D.B. (2013). Satisfaction with Massachusetts nursing home care was generally high during 2005-09, with some variability across facilities. Health Affairs. 32(8):1416-25.

Lin, J., Hsiao, C.T., Glen, R., Pai, J.Y., & Zeng, S.H. (2014). Perceived service quality, perceived value, overall satisfaction and happiness of outlook for long-term care institution residents. Health Expectations. 17(3):311-20.

National Committee for Quality Assurance (NCQA) (2016). HEDIS Measures.
http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures.aspx. Accessed March 2016.

Penchansky and Macnee, (1994). Initiation of medical malpractice suits: a conceptualization and test. Medical Care. 32(8): pp. 813–831

Reinhardt, J.P., Chichin, E., Posner, L., & Kassabian, S. (2014). Vital conversations with family in the nursing home: preparation for end-stage dementia care. Journal Of Social Work In End-Of-Life & Palliative Care. 10(2):112-26.

Van Uden, N., Van den Block, L., van der Steen, J.T., Onwuteaka-Philipsen, B.D., Vandervoort, A., Vander Stichele, R., & Deliens, L. (2013). Quality of dying of nursing home residents with dementia as judged by relatives. International Psychogeriatrics. 25(10):1697-707.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not Applicable.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator assesses the number of family or designated responsible party for long stay residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party members for long stay residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: Long-Stay Family questionnaire.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
While the frequency in which the questionnaires are administered is left up to the provider, they should at least be administered once a year. Once the questionnaire is administered to the family member or designated responsible party members for long stay residents, they have up to 2 months to return the questionnaire.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator includes all of the family or designated responsible party members for long stay residents that had an average response =>3 on the CoreQ: Long-Stay Family questionnaire.

We calculate the average satisfaction score for the individual family or designated responsible party member for long stay residents in the following manner:
- Respondents within the appropriate time window (see S.5) and who do not meet the exclusions (see S.11) are identified.
- A numeric score is associated with each response scale option on the CoreQ: Long-Stay Family questionnaire (that is, Poor=1, Average=2, Good=3, Very Good=4, and Excellent=5).
- The following formula is utilized to calculate the individual's average satisfaction score: [Numeric Score Question 1 + Numeric Score Question 2 + Numeric Score Question 3]/3
- The number of respondents whose average satisfaction score >=3 are summed together and function as the numerator.

For respondents with one missing data point (from the 3 items included in the questionnaire) imputation will be used (representing the average value from the other two available questions). For respondents with more than one missing data point, they will be excluded from the analyses (i.e., no imputation will be used for these family members). Imputation details are described further below (S.18).
No risk-adjustment is used (see S.13).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The target population is family or designated responsible party members of a resident residing in a SNF for at least 100 days. The denominator includes all of the individuals in the target population who respond to the CoreQ: Long-Stay Family questionnaire within the two month time window (see S.5) who do not meet the exclusion criteria (see S.10).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The denominator includes all of the family or the designated responsible party members for residents that have been in the SNF for 100 days or more regardless of payer status; who received the CoreQ: Long-Stay Family questionnaire (e.g. people meeting exclusions do not receive the questionnaire), and who responded to the questionnaire within the two month time window.

The length-of-stay (of the resident of the family member or designated responsible party) will be identified from MDS nursing facility records (MDS item A1600 “Entry Date”).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Please note, the resident representative for each current resident is initially eligible regardless of their being a family member or not. Only one primary contact per resident should be selected.

Exclusions made at the time of sample selection include: (1) family or designated responsible party for residents with hospice; (2) family or designated responsible party for residents with a legal court appointed guardian; (3) representatives of residents who have lived in the SNF for less than 100 days; and (4) representatives who reside in another country.

Additionally, once the survey is administered, the following exclusions are applied: a) surveys received outside of the time window (more than two months after the administration date) and b) surveys that have more than one questionnaire item missing.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Exclusions will be based on information from the Minimum Data Set (MDS) 3.0 assessment. Representatives of residents with the following criteria will be excluded:

(1) Residents on hospice. This is recorded in the MDS as Hospice O0100K1 = 1 (“the patient was on hospice in the last 14 days while not a resident”), O0100K2 = 1 (“the patient was on hospice in the last 14 days while a resident”), A1800=07 (“entered from hospice”), or A2100=07 (“discharged to hospice”).
(2) Residents with court appointed legal guardian for all decisions will be identified from nursing facility health information system.
(3) Residents who have lived in the SNF for less than 100 days will be identified from the MDS. This is recorded in the MDS (item A1600 “Entry Date”).
(4) Respondents who reside in another country, to be identified from nursing facility health information system.
(5) Respondents who have two or more missing data point are excluded from the analysis.
(6) Respondents that respond after the two month response period will be excluded.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
No stratification is used.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not Applicable.

S.16. **Type of score:**
Other (specify):
If other: Non-weighted score. Score is a percent.

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify the representatives of residents that have been residing in the SNF for 100 days or more. Length of stay so far is the MDS target date (TRGT_DT) - MDS admission date (A1900).

2. Take the representatives of residents that have been residing in the SNF for >=100 days and exclude the following:
a. Representatives of residents on hospice. This is recorded in the MDS as Hospice O0100K1 = 1 (“the patient was on hospice in the last 14 days while not a resident”), O0100K2 = 1 (“the patient was on hospice in the last 14 days while a resident”), A1800=07 (“entered from hospice”), or A2100=07 (“discharged to hospice”).
b. Residents with Court appointed legal guardian for all decisions as identified from nursing facility health information system.

3. Exclude representatives of residents who reside in another country.

4. Administer the CoreQ: Long-Stay Family questionnaire (See S.25) to the representatives that do not meet these exclusion criteria. Provide the family or designated responsible party member for the resident two months to respond to the survey.
a. Create a tracking sheet with the following columns:
  i. Date Administered
  ii. Date Response Received
  iii. Time to Receive Response: ([Date Response Received – Date Administered])
b. Exclude any surveys where Time to Receive Response >60 days (2 months)

5.Combine the CoreQ: Long-Stay Family questionnaire items to calculate a resident’ representative satisfaction score. Responses for each item should be given the following scores:
a.Poor = 1,
b.Average = 2,
c.Good = 3,
d.Very good =4 and
e.Excellent = 5.

6.Impute missing data if only one of the three questions are missing data. Drop all survey response if 2 or more survey questions have missing data.

7.Calculate resident's representative score from usable surveys.
a.Representative average score = (Score for Item 1 + Score for Item 2 + Score for Item 3) / 3.
b.Flag those representatives with a score equal to or greater than 3.0
i.For example, a representative of a resident rates their satisfaction on the three CoreQ questions as excellent = 5, very good = 4, and good = 3. The family member's total score will be 5 + 4 + 3 for a total of 12. The representative of the long-stay resident total score (12) will then be divided by the number of questions (3), which equals 4.0. Thus the representative's average satisfaction rating is 4.0. Since this person's average response is >3.0 they would be counted in the numerator. If it was <3.0 they would not be counted.

8.Calculate the facility's CoreQ: Long-Stay Family Measure which represents the percent of respondents with average scores of 3.0 or above.
a.CoreQ: Long-Stay Family Measure = ([number of respondents with an average score of =3.0] / [total number of valid responses])*100

9.No risk-adjustment is used.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
No sampling is used. No proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
1. Identify the representatives of residents that have been residing in the SNF for 100 days or more. This will be identified from MDS target date (TRGT_DT) - MDS admission date (A1900).

2.Take the representatives of residents that have been residing in the SNF for >=100 days and exclude the following:
a. Representatives of residents on hospice. This is recorded in the MDS as Hospice O0100K1 = 1 ("the patient was on hospice in the last 14 days while not a resident"), O0100K2 = 1 ("the patient was on hospice in the last 14 days while a resident"), A1800=07 ("entered from hospice"), or A2100=07 ("discharged to hospice").
b. Residents with Court appointed legal guardian for all decisions as identified from nursing facility health information system.

3. Exclude representatives of residents who reside in another country.
4. Administer the CoreQ: Long-Stay Family questionnaire to family or designated responsible party members for long-stay residents.

5. Instruct representatives that they must respond to the survey within 2 months.

6. The response rate for a center is calculated by counting the number of usable surveys returned divided by the number of surveys administered.
a. Surveys returned as undeliverable are not counted as usable.
b. Surveys with missing responses for two or more questions are also not counted as usable.
c. A minimum response rate of 30% needs to be achieved for results to be reported for a SNF.

7. Regardless of response rate, SNFs must also achieve a minimum number of 20 usable questionnaires (e.g. denominator). If after 2 months, less than 20 usable questionnaires are received than a facility level satisfaction measure cannot be reported.

8. All the questionnaires that are received (other than those that satisfy the exclusion criteria seen in section S.11) must be used in

the calculations.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data was uncommon in the CoreQ: Long Stay Family questionnaire testing (4.3% of any one of the 3 items). For representatives with one missing data point (from the 3 items included in the questionnaire) imputation will be used (representing the average value from the other questionnaire items). For family or designated response party members with more than one missing data point, they will be excluded from the analyses (i.e., no imputation will be used).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Healthcare Provider Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The collection instrument is the CoreQ: Long-Stay Family questionnaire and for exclusions the Resident Assessment Instrument Minimum Data Set (MDS) version 3.0 is used

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
CoreQ_Family_Testing_Final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: CoreQ: Long-Stay Family Measure
Date of Submission: 3/31/2016
Type of Measure:

☒Composite – ***STOP – use composite testing form***		☒ Outcome (*including PRO-PM*)
☐ Cost/resource		☐ Process
☐ Efficiency		☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly refle the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequenc of occurrence so that results are distorted without the exclusion; [12]
AND
If Family preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about Family preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on Family factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, b are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreemen with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testi hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure sc as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addres whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Family preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. Th substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of Familys who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference o $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may n demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims

☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: CoreQ: Long-Stay Family questionnaire	☒ other: CoreQ: Long-Stay Family questionnaire, Pilot CoreQ: Long-Stay Family questionnaire, Nursing Home Compare and CASPER

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
First, the Pilot CoreQ: Long-Stay Family questionnaire containing an extended list of questions included on the CoreQ: Long-Stay Family questionnaire was utilized for reliability and validity testing.

Second, data from the CoreQ: Long-Stay Family questionnaire was used to test the measure for reliability and validity.

Third, to validate the measure, we also utilized Certification and Survey Provider Enhanced Reporting (CASPER) Quality Indicators and data form Nursing Home Compare.

1.3. What are the dates of the data used in testing? Click here to enter date range
June, 2014-September, 2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: Individual Family

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
The testing and analysis included three data sources, one of which had additional variables collected for a subset of respondents:

1. The Pilot CoreQ: Long-Stay Family questionnaire was examined using responses from 1,324 Family members or resident representatives from a national sample of nursing facilities (Data Source #1).
    a. In addition, Family-level sociodemographic (SDS) variables were examined using this same sample of 1,324 Family members or resident representatives (#1 above) in nursing facilities across the US. (Data Source #1).
2. Validity testing of the Pilot CoreQ: Long-Stay Family questionnaire was examined using responses from 100 Family members or resident representatives from the Pittsburgh area. (Data Source #2).
3. CoreQ: Long-Stay Family measure was examined using 221 facilities and included responses from 6,192 Family members or resident representatives. These nursing facilities were located in multiple states across the US. (Data Source #3).

Some basic descriptive characteristics of these facilities (data sources) are provided below.

Table 1.5: Descriptive Statistics of Centers Included in the Analysis

Data Source	Average Number of Licensed Beds	Average Daily Census	Sample Size of Family members (N)
Listed #1 (above)	136	122	1,324
Listed #2 (above)	202	188	100
Listed #3 (above)	142	131	6,192

1.6. How many and which Family members were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of Family included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how Family were selected for inclusion in the sample*)

Family Level of Analysis

Data was used from the CoreQ: Long-Stay Family questionnaire. The questionnaire was administered to all eligible long-stay family (with the exclusions described in the Specification part of this application). The testing and analysis included:

1. The Pilot CoreQ: Long-Stay Family questionnaire was examined using responses from 1,324 family members or resident representatives from a national sample of nursing facilities. (Data #1)
    a. In addition, Family-level sociodemographic (SDS) variables were examined using this same sample of 1,324 family members (Data #1 above) in nursing facilities across the US.
2. Validity testing of the Pilot CoreQ: Long-Stay Family questionnaire was examined using responses from 100 family members from the Pittsburgh area. (Data #2)
3. CoreQ: Long-Stay Family questionnaire MEASURE was examined using 221 facilities and included responses from 6,192 family members or resident representatives. These nursing facilities were located in multiple states across the US. (Data #3)

The descriptive characteristics of the family members are given in the following table that includes information from all of the data used (the education level and race information comes only from the sample described above with 1,324 respondents, as this data was not collected for the other samples).

Table 1.6: Respondent Demographics (all samples pooled)

DEMOGRAPHICS		Percent
Are you male or female?	Male	30%
	Female	70%
What year were you born?	Average	1946
What is the highest grade or level of school that you have completed?	Some HS	7%
	HS or GED	32%
	Some College/ 2yr Degree	32%
	4yr College Degree	15%
	>4yr College Degree	15%
What is your race?	White	92%
	Black	7%
	Asian	1%
	Native Hawaiian	0%
	American Indian	0%

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

We conducted two levels of testing in the development of the CoreQ: Long-Stay Family measure. The first focused on testing (e.g., reliability, validity, and exclusions) of the CoreQ: Long-Stay Family questionnaire. The first source of data (pilot data) was utilized in developing and choosing the items to be included in the CoreQ: Long-Stay Family questionnaire. This included using a questionnaire with 18 items. Below we call this the Pilot CoreQ: Long-Stay Family questionnaire (i.e., Data #1, above). A subset of 100 family members from Data #1 was chosen in Data #2 to conduct a lagged re-administration of the same survey to measure agreement in response for the same family members regarding care the same period of time.

Once the CoreQ: Long-Stay Family questionnaire was developed, a second source of data was used to test the validity of the CoreQ: Long-Stay Family measure (i.e., facility and summary score validity). This second data source is described above (i.e.221 facilities including responses from 6,192 family members [Data #3, above]).

1.8 What were the Family-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, Family-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each Family (e.g. census tract), or Family community characteristics (e.g. percent vacant housing, crime rate).

The following Family-level sociodemographic variables were available for analysis. For the distributions of these categories, see Tables 1.6 above.

- Age
  - Exact date of birth
- Sex
  - Male
  - Female

- Highest level of education
  - Some high school, but did not graduate
  - High school graduate or GED
  - Some college or 2 year degree
  - 4 year college graduate
  - More than 4 year college degree
- Race
  - White
  - Black or African American
  - Asian
  - Native Hawaiian or other Pacific Islander
  - American Indian or Alaskan Native.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

We measured reliability at the: (1) data element level; (2) the person/questionnaire level; and, (3) at the measure (i.e., facility) level. More detail of each analysis follows.

(1) DATA ELEMENT LEVEL

To determine if the CoreQ: Long-Stay Family questionnaire items were repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, we re-administered the questionnaire to family members 1 month after their first survey. The Pilot CoreQ: Long-Stay Family questionnaire had responses from 100 family members; we re-administered the survey to 50 of these same family members. The re-administered sample was a sample of convenience as they represented family members from the Pittsburgh area (the location of the team testing the questionnaire). To measure the agreement, we calculated first the distribution of responses by question in the original round of surveys, and then again in the follow-up surveys (they should be distributed similarly); and second, calculated the correlations between the original and follow-up responses by question (they should be highly correlated).

(2) PERSON/QUESTIONNAIRE LEVEL

Having tested whether the data elements matched between the pilot responses and the re-administered responses, we then examined whether the person-level results matched between the Pilot CoreQ: Long-Stay Family questionnaire responses and their corresponding re-administered responses. In particular, we calculated the percent of time that there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re-administered response was poor, average, good, very good or excellent.

(3) MEASURE (FACILITY) LEVEL

Last, we measured stability of the facility-level measure when the facility's score is calculated using multiple "draws" from the same population. This measures how stable the facility's score would be if the underlying family members are from the same population but are subject to the kind of natural sample variation that occurs over time. We did this by bootstrap with 10,000 repetitions of the facility score calculation, and present the percent of facility resamples where the facility score is within 1 percentage point, 3 percentage points, 5 percentage points, and 10 percentage points of the original score calculated on the Pilot CoreQ: Long-Stay Family questionnaire sample.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

(1) DATA ELEMENT LEVEL

Table 2a2.3.a shows the three CoreQ: Long-Stay Family questionnaire items, and the response per item for both the pilot survey of 100 family members and the re-administered survey of 50 family members. The responses in the pilot survey are not statistically significant from the re-administered survey. This shows that the data elements were highly repeatable and produced the same results a high proportion of the time when assessing the same population in the same time period.

Table 2a2.3.a: CoreQ: Long-Stay Family Questionnaire Responses from the Pilot and Re-administered Survey

Questionnaire Item	Response	Percent [Pilot Survey (N=100)]	Percent [Re-Administered Survey (N=50)]
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	4.5%	4%
	Average	14%	13%
	Good	24%	25%

	Very Good	35%	36%
	Excellent	20%	19%
2. Overall, how would you rate the staff?	Poor	2%	3%
	Average	12%	11%
	Good	22%	22%
	Very Good	34%	32%
	Excellent	23%	22%
3. How would you rate the care you receive?	Poor	3%	3%
	Average	14%	13%
	Good	22%	22%
	Very Good	33%	31%
	Excellent	21%	22%

NO SIGNIFICANT DIFFERENCES AT p=0.01

Table 2a2.3.b shows the average of the percent agreement from the first survey score to the second survey score for each item in the CoreQ: Long-Stay Family questionnaire. This shows very high levels of agreement.

Table 2a2.3.b: Average Percent Agreement between the Pilot and Re-administered Surveys

Questionnaire Item	Percent Agreement
4. In recommending this facility to your friends and family, how would you rate it overall?	97.1%
5. Overall, how would you rate the staff?	98.8%
6. How would you rate the care your family member received?	97.5%

(2) PERSON/QUESTIONNAIRE LEVEL

Table 2a2.3.c shows the CoreQ: Long-Stay Family questionnaire items, and the agreement in response per item for both the pilot survey of 100 family members compared with the re-administered survey of 50 family members. The person-level responses in the pilot survey are not statistically significant from the re-administered survey. This shows that a high percent of time there was agreement between whether or not the pilot response was poor, average, good, very good or excellent, and whether or not the re-administered response was poor, average, good, very good or excellent. Table 2a2.3.d shows the agreement between the pilot and re-administered responses. In summary, 97% or more of the re-administered responses agreed with their corresponding pilot responses, in terms of whether or not they were rated in the categories of poor or average or good, very good or excellent.

Table 2a2.3.c: Average Percent Agreement between Responses per Item for the Pilot Survey and Re-Administered Survey

Questionnaire Item	Response	Percent Person-Level Agreement in Response for the Pilot Survey (N=100) vs. Re-Administered Survey (N=50)
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	98%
	Average	97%
	Good	97%
	Very Good	98%
	Excellent	97%
2. Overall, how would	Poor	98%

you rate the staff?	Average	96%
	Good	98%
	Very Good	99%
	Excellent	99%
3. How would you rate the care you receive?	Poor	99%
	Average	99%
	Good	97%
	Very Good	98%
	Excellent	97%

Table 2a2.3.d: Average Percent Agreement between Response Options for the Pilot Survey and Re-Administered Survey

		Re-Administered Response	
		Poor (1) or Average (2)	Good (3), Very Good (4), or Excellent (5)
Pilot Response	Poor (1) or Average (2)	98.5%	98.8%
	Good (3), Very Good (4), or Excellent (5)	98.5%	98.7%

(3) MEASURE (FACILITY) LEVEL

After having performed the 10,000-repetition bootstrap, 11.5% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample, 20.9% were within 3 percentage points, 30.4% were within 5 percentage points, and 42.2% were within 10 percentage points.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
In summary, the measure displays a high degree of element-level, questionnaire-level, and measure (facility)-level reliability. First, the CoreQ: Long-Stay Family questionnaire data elements were highly repeatable, with pilot and re-administered responses agreeing between 97% and 99% of the time depending on the question. That is, this produced the same results a high proportion of the time when assessed in the same population in the same time period. Second, the questionnaire level scores were also highly repeatable, with pilot and re-administered responses agreeing 98% of the time (or more). Third, a facility drawing family members from the same underlying population will only vary modestly. The 10,000-repetition bootstrap results show that the CoreQ: Long-Stay Family measure scores from the same facility are moderately stable given the minimum sample size of 20 was set for this measure; and the maximum sample size was 95.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☐ **Empirical validity testing**
☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
In the development of the CoreQ: Long-Stay Family questionnaire, three sources of data were used to perform three levels of validity testing. These are described above in Section 1.5.

The first source of data (data from a sample of convenience collected near the researchers developing the questionnaire in Pittsburgh) was used in developing and choosing the format to be utilized in the CoreQ: Long-Stay Family questionnaire (i.e., response scale).

The second source of data, was pilot data collected from a national sample of 1,324 family members. This data was used in choosing the items to be used in the CoreQ: Long-Stay Family questionnaire (i.e., questionnaire items). This data was also used in examining Family-level sociodemographic (SDS) variables.

The third source of data (collected from 221 facilities) was used examine the validity of the CoreQ: Long-Stay Family measure (i.e., facility and summary score validity). These family members / nursing facilities were from multiple states across the U.S.

Thus, the following sections describe this validity testing:
1. Validity Testing of the questionnaire format used in the CoreQ: Long-Stay Family questionnaire (using data source 1, from above);
2. Testing the items for the CoreQ: Long-Stay Family questionnaire (using data source 2, from above);
3. Testing to determine if a sub-set of items could reliably be used to produce an overall indicator of satisfaction (Core Q: Long-Stay Family measure) (using data source 3, from above);
4. Validity testing for the CoreQ: Long-Stay Family measure (also using data source 1, from above).

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Family Questionnaire

A. The face validity of the domains used in the CoreQ: Long-Stay Family questionnaire was evaluated via a literature review. The literature review was conducted to examine important areas of satisfaction for LTC family. Specifically, the research team examined 12 commonly used satisfaction surveys and reports to determine the most valued domains when looking at satisfaction. These surveys were identified by completing internet searches in PubMed and Google. Key terms that were searched included: Family satisfaction, long-term care satisfaction, and elderly satisfaction.

B. The face validity of the domains was also examined using a focus group of family members. The overall ranking used was 1=Most important and 22=Least important. That is family members were asked to rank the domains from most important to least important. The respondents were family members (N=40) of residents in five nursing facilities in the Pittsburgh region.

C. The face validity of the Pilot CoreQ: Long-Stay Family questionnaire response scale was also examined. The respondents were family members (N=40) with residents in five nursing facilities in the Pittsburgh region The percent of respondents that stated they "fully understood" how the response scale worked, could complete the scale, AND in cognitive testing understood the scale was used.

D. The Flesch-Kinkaid scale was used to determine if respondent correctly understood the questions being asked (Streiner & Norman, 1995).

Streiner, D. L. & Norman, G.R. (1995). Health measurement scales: A practical guide to their development and use. 2nd ed. New York: Oxford.

2. Testing the Items for the CoreQ: Long-Stay Family Questionnaire

The second series of validity testing was used to further identify items that should be included in the CoreQ: Long-Stay Family questionnaire. This analysis was important, as all items in a satisfaction measure should have adequate psychometric properties (such as low basement or ceiling effects). For this testing, (1) A pilot group of 40 family members was first used in focus groups; (2) a Pilot version of the CoreQ: Long-Stay Family questionnaire survey was administered consisting of 18 items (N= 1,324 family members). The testing consisted of:
A. Family members were asked to rate the 18 different satisfaction questions related to their experience in SNFs. This was conducted with a pilot group of 40 family members in focus groups.

B. The Pilot CoreQ: Long-Stay Family questionnaire items performance with respect to the distribution of the response scale and with respect to missing responses. (Using 1,324 family members described above)
C. The intent of the Pilot instrument was to have items that represented the most important areas of satisfaction (as identified above) in a parsimonious manner. Additional analyses such as exploratory factor analysis (EFA) were used to eliminate items in the Pilot instrument. This was an iterative process that included using Eigenvalues from the principal factors (unrotated) and correlation analysis of the individual items. (using 1,324 family members described above)

3. To determine if a Sub-Set of Items could be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Family Measure).

The CoreQ: Long-Stay Family measure under development was meant to represent overall satisfaction with as few items as possible. The testing given below describes how this was achieved.

A. To support the construct validity that the idea that the CoreQ items measured a single concept of "satisfaction" – we performed a correlation analysis using all items in the instrument.

B. **B.** In addition, using all items in the instruments a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item was examined.

4. Validity Testing for the Core Q: Long-Stay Family Measure.

A. To determine if the 3 items in the CoreQ: Long-Stay Family questionnaire were a reliable indicator of satisfaction, the correlation between these three items (the "CoreQ: Long-Stay Family Measure") and ALL of the items on the Pilot CoreQ instrument was conducted.

B. We performed additional validity testing of the facility-level CoreQ: Long-Stay Family measure by examining the correlations between the CoreQ: Long-Stay Family measure scores and i) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, and ii) several other quality metrics from Nursing Home Compare. If the CoreQ Long Stay Family scores correlate negatively with the measures that decrease as they get better, and positively with the measures that increase as they get better, then this supports the validity of the CoreQ Long Stay Family measure.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Family Questionnaire

A. The face validity of the domains used in the CoreQ: Long-Stay Family questionnaire was evaluated via a literature review (described above).

The research team examined the surveys and reports to identify the different domains that were included. The research team scored the domains by simply counting if an instrument included the domain. Table 2b2.3.a gives the domains that were found throughout the search, as well as a score. An example is the domain clinical care, this was used in 10 out of the 12 surveys identified in the literature. An interpretation of this finding would be that items addressing clinical care are extremely important in satisfaction surveys. These domains were used in developing the pilot CoreQ: Long-Stay Family questionnaire items.

Table 2b2.3.a: Survey Domain Score out of 12

Domain	Score out of 12	Domain	Score out of 12
Food	11	Spiritual	4
Activities	10	Confidence in Caregivers	3
Administration	10	Language and Communication	3
Clinical Care	10	Personal Suite	3
Staff Interaction	10	Therapy	3
Choice and Decision Making	9	Care Access	2

Facility Environment	9	Case Manager	2
Security and Safety	9	Comfort	2
Overall	8	Maintenance	2
Staff Overall	7	Move In	2
Autonomy and Privacy	6	Non-Clinical Staff Services	2
Housekeeping	6	Transitions	2
Personal Care	6	Transportation	2
Recommend facility	6	Emergency Response	1
Resident to Resident Friendships	5	Finances	1
Family Involvement	4	Time	1
Resident to Staff Friendships	4	Trust	1

B. The face validity of the domains was also examined using family members. The following abbreviated table shows the rank of importance for each group of domains. The overall ranking used was 1=Most important and 22=Least important. The ranking of the 3 areas used in the CoreQ: Long-Stay Family questionnaire are shown. Note, the food domain was ranked third – but was excluded from the CORE Q based on additional analyses showing that it was highly correlated with the overall domain; thus, it added little to the measure.

Table 2b2.3.b: Face Validity Abbreviated Results

Domain / Question	Average Rank
OVERALL (In recommending this facility to your friends and family, how would you rate it overall?)	4
STAFF (Overall, how would you rate the staff?)	1
CARE (How would you rate the care you receive?)	2

C. The face validity of the pilot CoreQ: Long-Stay Family questionnaire response scale was also examined. Table 2b2.3.c gives the percent of respondents that stated they "fully understood" how the response scale worked, could complete the scale, AND in cognitive testing understood the scale.

Table 2b2.3.c: Respondent's Understanding of Response Scale

Scale Format	Residents /Family
Yes – No	100%
Yes – Somewhat – No	100%
Always – Usually – Sometimes –Never	100%
Very happy – Somewhat happy – Unhappy	100%
Excellent – Good – Fair – Poor	100%
Very Good – Good – Average – Poor – Very Poor	100%
Very Satisfied – Satisfied – Neither Satisfied or Dissatisfied – Dissatisfied – Very Dissatisfied	100%
4 Point Satisfaction Scale (1=Very unsatisfied, 2=Unsatisfied, 3=Neutral, 4=Satisfied)	100%
5 Point Likert Scale (1=Poor, 2=Average, 3=Good, 4=Very Good, 5=Excellent)	100%
Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree	95%
5 Point Importance Scale (1=Very important, 5=Very unimportant)	95%

5 Point Expectancy Scale (1=Not met, 2=Nearly met, 3=Met, 4=Exceeded, 5=Far exceeded expectations)	90%
10 Point Satisfaction Scale (1=Poor, 10=Excellent)	90%
8 Point Satisfaction Scale (1=Very dissatisfied, 2=Dissatisfied, 3=Somewhat dissatisfied, 4=Neither satisfied nor dissatisfied, 5=Somewhat satisfied, 6=Satisfied, 7=Very satisfied, 8=No response)	85%

Note: Highlighted cell represents the scale used in the CoreQ.

D. The CoreQ: Long-Stay Family questionnaire was purposefully written using simple language. No a priori goal for reading level was set, however a Flesch-Kinkaid scale score of six, or lower, is achieved for all questions.

2. Testing the Items for the CoreQ: Long-Stay Family Questionnaire

A. Each family member was asked to rate on a scale of 1 to 10 (with 10 as the best) how important they thought the question was for evaluating the experience with SNF care. The three questions included in the COREQ were highly rated out of all the questions and in analysis of family member's responses to 18 questions. That is, these three items were shown to provide unique information to distinguish satisfaction with SNFs. Specifically, "In recommending this facility to your friends and family, how would you rate it overall?" had an average score of 9.69; "Overall, how would you rate the staff?" had an average score of 9.6; and, "How would you rate the care you receive?" had an average score of 9.5. This shows a very pervasive influence of the satisfaction items with the experience of SNF care. See Table 1c.5 (Appendix)

B. The pilot CoreQ: Long-Stay Family questionnaire items are shown in Table 2b2.3.d. This shows that the items performed well with respect to the distribution of the response scale and with respect to missing responses.

C. Using all items in the instruments (excluding the global item Q1 ("How would you rate the facility?")) exploratory factor analysis (EFA) was used to evaluate the construct validity of the measure. The Eigenvalues from the principal factors (unrotated) are presented in the Table below. In this analysis, the first Eigenvalue is overwhelmingly greater than the second Eigenvalue, this supports the proposition that the CoreQ instrument is measuring a single global concept of customer satisfaction – rather than a number of sub-concepts of customer satisfaction. Sensitivity analyses using principal factors and rotating provide highly similar findings.

Table 2b2.3.e: Exploratory Factor Analysis Results

	Long-Stay Family
Factor 1	11.73
Factor 2	0.61

3. To determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Family measure).

A. To support the construct validity that the idea that the CoreQ items measured a single concept of "satisfaction" – we performed a correlation analysis using all items in the instrument. The analysis identifies the pairs of CoreQ items with the highest correlations. The highest correlations are shown in the Table 2b2.3.f. Items with the highest correlation are potentially providing similar satisfaction information. Because items with the highest correlation were potentially providing similar satisfaction information they could be eliminated from the instrument. Note, the table provides 7 sets of correlations, however the analysis was conducted examining all possible correlations between items.

Table 2b2.3.f: CoreQ: Long-Stay Family Questionnaire Example Item Correlations

	Family
Highest Correlation	Q1-Q10 (.845)
Next highest Correlation	Q1-Q2 (.841)
Next highest Correlation	Q1-Q6 (.826)
Next highest Correlation	Q1-Q5 (.757)
Next highest Correlation	Q1-Q9 (.782)
Next highest Correlation	Q1-Q18 (.710)

RESULT = ITEMS TO DROP

B. In addition, using all items in the instrument a factor analysis was conducted. Using the global items Q1 ("How satisfied are you with the facility?") the Cronbach's Alpha of adding the "best" additional item is shown in the table below. Cronbach's alpha measures the internal consistency of the values entered into the factor analysis; a value of 0.7 or higher is generally considered acceptably high. The additional item(s) is considered best in the sense that it is most highly correlated with the existing item, and therefore provides little additional information about the same construct. So this analysis was also used to eliminate items. Note, table 2b2.3.g again provides 7 sets of correlations, however the analysis was conducted examining all possible correlations between items.

Table 2b2.3.g: Secondary Correlation Analysis of CoreQ: Long-Stay Family Questionnaire Items

	Family
Q1 + last satisfaction item ADD	Q10 (.943) Q6 (.939) Q2 (.935)
Q1 + ADD ADD	Q2 + Q6 (.931) Q10 + Q6 (.931) Q2 + Q10 (.929)
Q1 + ADD ADD	Q10 + Q6 (.939) Q9 + Q6 (.935) Q2 +Q6 (.935)

Thus, using the correlation information and factor analysis 3 items representing the CoreQ: Long-Stay Family questionnaire were identified.

4. ***Validity Testing for the Core Q: Long-Stay Family Measure.***

The overall intent of the analyses described above was to identify if a sub-set of items could reliably be used to produce an overall indicator of satisfaction, the CoreQ: Long-Stay Family questionnaire.

A. The items were all scored according to the rules identified elsewhere. The same scoring was used in creating the 3 item CoreQ: Long-Stay Family questionnaire summary score and the satisfaction score using the Pilot CoreQ: Long-Stay Family questionnaire. The correlation was identified as having a value of 0.90.

That is, the correlation score between actual the "CoreQ: Long-Stay Family Measure" and all of the 18 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 3 items (much less burdensome, and therefore likely to yield a higher response rate) or the 18 item Pilot instrument. Thus, we only included the three measures as additional measures did not provide additional information for a quality measure to assess a facilities satisfaction score. Additional questions may help with quality improvement efforts to identify specific areas of satisfaction or dissatisfaction.

B. We performed additional validity testing of the facility-level CoreQ: Long-Stay Family measure by measuring the correlations between the CoreQ: Long-Stay Family measure scores and A) measures of regulatory compliance and other quality metrics from the Certification and Survey Provider Enhanced Reporting (CASPER) data, B) several other quality metrics from Nursing Home Compare, C) risk-adjusted discharge to community measure [NQF# 2858], and D) risk adjusted PointRight® Pro 30™ Rehospitalizations [NQF# 2375].

CoreQ: Long-Stay Family measure is the percentage of family members of residents discharged from the facility within 100 days of admission from a hospital to the nursing facility who, on average for the three CoreQ items included in the measure, rated the facility >= 3. We measured satisfaction using family's responses to the three items from the CoreQ: Long-Stay Family questionnaire (see Table 2a2.3.a).

The summary score from the 3 CoreQ: Long-Stay Family questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the Family is calculated. The facility score represents the percent of family members with average scores of 3 or above. This score should be associated with quality. Therefore, for each facility in the sample the correlation with other quality indicators was examined.

(v) Relationship with CASPER Quality Indicators

Certification and Survey Provider Enhanced Reporting (CASPER) contains data collected as part of state/federal nursing home inspections. In short, nursing facilities that accept residents with Medicare and/or Medicaid payments are surveyed. This includes most (i.e., 97% [16,000 facilities]) nursing homes in the U.S. The survey process occurs approximately yearly, and includes the recording of many quality characteristics of the nursing home. These include restraint use; pressure ulcers; catheter use; antipsychotic use; antidepressant use; antianxiety use; and, use of hypnotics. These are commonly used quality indicators used for examining the quality of nursing homes.
In addition, when a nursing home is determined not to meet a certification minimum standard a deficiency citation is issued. These deficiency citations are also commonly used in the analyses of the quality of nursing homes. Approximately 180 deficiency citations exist and are grouped into 16 categories. These 16 categories group like areas together. They were developed by CMS and have considerable face validity; although, one limitation of using these categories is that they were not defined using empirical estimation (such as factor analysis). One category groups together 25 "quality of care" deficiency citations. In addition, for all deficiency citations a determination of the scope and severity of the problem(s) identified is also made. One of 12 categories is used which are labeled "A" through "L," with L having the highest severity and scope. The most severe (i.e., JKL) are used in this analysis. Thus, we would expect a negative correlation between family satisfaction and the number and severity of deficiencies cited by the State Survey agency.

Table 2b2.3.g: Correlation results between the CoreQ Long Stay Family Questionnaire Measure Score and CASPER Quality Indicators

Quality Indicator	Correlation with Satisfaction Summary Score	P-Value
Restraint Use	-0.28	<0.01
Pressure Ulcers	-0.04	0.51
Catheter Use	-0.03	0.70
Antipsychotic Use	-0.14	0.04
Antidepressant Use	0.08	0.23
Antianxiety Use	-0.09	0.19
Use of Hypnotics	-0.10	0.16
Deficiency Citation	-0.08	0.23

(vi) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels

Nursing Home Compare (NHC) is a nursing home report card. After several years of pilot testing, the Centers for Medicare and Medicaid Services (CMS) released this report card on the world-wide web in November of 2002. Briefly, Nursing Home Compare provides information for facility location, structural factors (such as ownership), and staffing characteristics (such as registered nurse [RN] staffing levels). Most significantly, standardized quality information is presented in what are called Quality Measures (QMs). These are calculated from MDS information.

At the time period of for this study (i.e., 2014) CMS reported on 19 measures – these are called the core Quality Measures. The Quality Measures address specific areas of resident care, 5 are for short-stay residents and 14 are for long-stay residents. Long-stay measures are for those residents staying at a facility 3 months or more and short-stay measures are for residents staying at a facility less than 3 months. The long-stay measures are most pertinent to the CoreQ: Long-Stay Family questionnaire; therefore, these were used in the analyses.

Nursing Home Compare also uses a five-star rating for facilities. This is based on information from the health inspection, direct care staffing, and the MDS quality measures. A five star facility is the highest score and a 1 star facility the lowest score. With respect to staffing, two measures are used: 1) RN hours per Family day; and 2) total staffing hours (RN+ LPN+ nurse aide hours) per Family day.

Table 2b2.3.h: Correlation results between the CoreQ Long Stay Family Questionnaire Measure Score and NHC Quality Indicators, Five Star ratings, and staffing levels

Quality Indicator	Correlation with Satisfaction Summary Score	P-Value

	MEASURE	
Percent of long-stay residents experiencing one or more falls with major injury.	-0.17	0.01
Percent of long-stay residents with a urinary tract infection	-0.29	0.09
Percent of long-stay residents who self-report moderate to severe pain	-0.24	0.15
Percent of long-stay high-risk residents with pressure ulcers	-0.21	0.22
Percent of long-stay low-risk residents who lose control of their bowels or bladder	-0.11	0.01
Percent of long-stay residents who have/had a catheter inserted and left in their bladder	-0.32	0.07
Percent of long-stay residents who were physically restrained	-0.41	0.09
Quality Indicator	**Correlation with Satisfaction Summary Score MEASURE**	**P-Value**
Percent of long-stay residents whose need for help with daily activities has increased	-0.33	0.03
Percent of long-stay residents who lose too much weight	-0.19	0.21
Percent of long-stay residents who have depressive symptoms	-0.13	0.10
Percent of long-stay residents assessed and given, appropriately, the seasonal influenza vaccine	0.40	0.08
Percent of long-stay residents assessed and given, appropriately, the pneumococcal vaccine	0.30	0.09
Percent of long-stay residents who are administered antipsychotic medications	0.16	0.10
Five-Star rating	0.32	0.13
RN hours per resident day	0.45	0.10
Total staffing hours	0.42	0.05

(vii) Relationship with the risk-adjusted Discharge to Community Measure

The Discharge to Community measure [NQF# 2858] determines the percentage of all new admissions from a hospital who are discharged back to the community within 100 days and remain out of any skilled nursing center for the next 30 days. The measure, referring to a rolling year of MDS entries, is calculated each quarter and includes all new admissions to a SNF regardless of payor source. Unsuccessful discharges will result in the resident becoming a long stay resident, which we hypothesize would increase family member dissatisfaction in SNFs with poor discharge to community rates.

The results of testing for correlation between Risk-adjusted discharge to community measure (from 2015q1) and the CoreQ: Long-Stay Family questionnaire are provided in the table below.

Table 2b2.3.i: Correlation results between the CoreQ Long Stay Family Measure and Risk-adjusted Discharge to Community Measure

CoreQ: Long-Stay Family	Correlation with Risk-adjusted discharge to community measure	P-Value
Q1: In recommending this facility to your	-0.03	0.65

friends and family, how would you rate it overall?		
Q2: Overall, how would you rate the staff?	-0.06	0.36
Q3: How would you rate the care you family member received?	-0.05	0.44
CoreQ: Long-Stay Family summary score	-0.05	0.48

(viii) Relationship with the risk adjusted PointRight® Pro 30™ Rehospitalizations

PointRight® Pro 30™ [NQF# 2375] is an all-cause, risk adjusted rehospitalization measure. It provides the rate at which all patients (regardless of payer status or diagnosis) who enter skilled nursing facilities from acute hospitals and are subsequently rehospitalized during their SNF stay, within 30 days from their admission to the SNF. Individuals who are rehospitalized after admission are much more likely to become a long stay residents. We hypothesize family members would therefore be more dissatisfied on average in SNFs with high short stay resident rehospitalization rates.

The results of testing for correlation between Risk-adjusted PointRight® Pro 30™ Rehospitalizations measure (from 2015q2) and the CoreQ: Long-Stay Family questionnaire are provided in the table below.

Table 2b2.3.j: Correlation results between the CoreQ Long Stay Family Measure and Risk-adjusted PointRight® Pro 30™ Rehospitalizations Measure

CoreQ: Long-Stay Family	Correlation with Risk-adjusted PointRight® Pro 30™ Rehospitalizations measure	P-Value
Q1: In recommending this facility to your friends and family, how would you rate it overall?	-0.21	<0.01
Q2: Overall, how would you rate the staff?	-0.18	<0.01
Q3: How would you rate the care you family member received?	-0.20	<0.01
CoreQ: Long-Stay Family summary score	-0.21	<0.01

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: Long-Stay Family Questionnaire

A. The literature review shows that domains used in the Pilot CoreQ: Long-Stay Family questionnaire items have a high degree of both face validity and content validity.

B. Family's overall rankings, show the general "domain" areas used indicates a high degree of both face validity and content validity.

C. The results show that 100% of Family's are able to complete the response format used. This testing indicates a high degree of both face validity and content validity.

D. The Flesch-Kinkaid scale score achieved for all questions indicates that respondents have a high degree of understanding of the item.

2. ***Testing the Items for the CoreQ: Long-Stay Family Questionnaire***

A. The percent of missing responses for the items is very low. The distribution of the summary score is wide. This is important for quality improvement purposes, as nursing facilities can use benchmarks etc.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

3. Testing to Determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The Core Q: Long-Stay Family measure)

A. Using the correlation information of the Core Q: Long-Stay Family questionnaire (18 items) and the 3 items representing the CoreQ: Long-Stay Family questionnaire a high degree of correlation was identified. This testing indicates a high degree of criterion validity.

B. EFA shows that one factor explains the common variance of the items. A single factor can be interpreted as the only "concept" being measured by those variables. This means that the instrument measures the global concept of satisfaction and not multiple areas of satisfaction. This supports the validity of the CoreQ instrument as measuring a single concept of "customer satisfaction". This testing indicates a high degree of criterion validity.

4. Validity Testing for the Core Q: Long-Stay Family Measure

A. The correlation of the 3 item CoreQ: Long-Stay Family measure summary score (identified elsewhere in this document) with the overall satisfaction score (scored using all data and the same scoring metric) gave a value of 0.90.

That is, the correlation score between actual the "CoreQ: Long-Stay Family Measure" and all of the 18 items used in the Pilot instrument indicates that the satisfaction information is approximately the same if we had included either the 3 items or the 18 item Pilot questions.

This indicates that the CoreQ: Long-Stay Family measure score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B.

(ix) Relationship with CASPER Quality Indicators

The CASPER Quality Indicators all had negative correlation with the CoreQ: Long-Stay Family measure as expected (higher satisfaction is associated with better quality). These correlations range from ± 0.03 to 0.28. The CoreQ: Long-Stay Family measure is associated with these quality indicators. This testing indicates a reasonable degree of construct validity and convergent validity.

(x) Relationship with Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels

The Nursing Home Compare (NHC) Quality Indicators, Five Star ratings, and staffing levels had a moderate to high level of correlation with the CoreQ: Long-Stay Family measure. These correlations range from ± 0.11 to 0.45. The CoreQ: Long-Stay Family measure is associated with these quality indicators, and always in the hypothesized direction (good correlates with good). In particular, as emphasized in the structure-process-outcome framework of the evidence section, the link between staffing and customer satisfaction is particularly high, as confirmed by the correlation coefficients 0.45 for RN hours per resident-day and 0.42 for total staffing hours per resident day. This testing indicates a reasonable degree of construct validity and convergent validity.

(xi) Relationship with the risk-adjusted Discharge to Community Measure

The risk-adjusted Discharge to community measure was negatively correlated to the CoreQ: Long-Stay Family measure. The correlations range from -0.03 to -0.06, all of which are not statistically significant at the p-value

of 0.05. This was not as hypothesized which may be related to some SNFs that specialize in long stay, have very low discharge to community rates as admissions do not have a plan to go home.

(xii) Relationship with the risk adjusted PointRight® Pro 30™ Rehospitalizations

The risk-adjusted PointRight® Pro 30™ Rehospitalizations was negatively correlated to the CoreQ: Long-Stay Family measure. The correlations range from -0.18 to -0.21, and all of them were statistically significant at the p-value of 0.05. This is expected because lower rehospitalization rates (an indicator of high quality) are associated with higher satisfaction scores. This was as hypothesized. This testing indicates a reasonable degree of construct validity and convergent validity.

As noted by Mor and associates (2003, p.41) "there is only a low level of correlation among the various measures of quality" In long term care settings. Castle and Ferguson (2010) also show the pattern of findings of quality indicators in nursing facilities is consistently moderate with respect to the correlations identified. The magnitude of correlations of the CoreQ with quality metrics are consistent with these findings in this setting.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
To develop the CoreQ: Long-Stay Family measure, we convened an expert panel to advise us on aspects such as which exclusions to apply to the measure, with the goal to make sure as many family members who are capable of giving a response are included as possible, and that the voice of the Family is included not proxies.

The exclusion analysis included 221 nursing homes that have used the CoreQ: Long-Stay Family measure. These facilities were included in multiple states across the US (this is data source 3, from above).

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
The expert panel advised us to exclude: 1) Family members of residents receiving hospice care; and (2) Family members of residents with a legal court appointed guardian.

In addition we exclude; (3) Family members of residents who have lived in the SNF for less than 100 days; (4) Respondents who have one or more missing data point (on the COREQ items); and (5) surveys received outside of the time window (more than two months after the administration date).

These exclusions are often used with satisfaction surveys (Sangl et al., 2007). The exclusions were made at the time of data collection, so we are able to report descriptive statistics regarding the number of exclusions made.

The exclusion analysis included responses from 221 facilities (described elsewhere). The exclusions were tracked and from these facilities included 2% Family members of residents with hospice; and 4% family members with a legal court appointed guardian.

Sangl, J., Bernard, S., Buchanan, J., Keller, S., Mitchell, N., Castle, N.G., Cosenza, C., Brown, J., Sekscenski, E., and Larwood, D. (2007). The development of a CAHPS instrument for nursing home Familys. Journal of Aging and Social Policy, 19(2), 63-82.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If Family preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
These exclusions were applied because such family members were either unable to provide an independent response or for whom the burden of completing a questionnaire is inappropriate given their residents clinical situation (e.g. hospice residents who are extremely sick and in the dying process). Therefore, the value of excluding these respondents takes into account burden on respondents and likely distortion of the results.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in Family characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

No research to date has risk adjusted or stratified satisfaction information from nursing facilities. Testing on this was conducted as part of the development of the federal initiative to develop a CAHPS®[2] Nursing Home Survey to measure nursing home residents' experience (hereafter referred to as NHCAHPS). No empirical, theoretical or stratified reporting of satisfaction information was recommended as the evidence showed that no clear relationship existed with respect to family characteristics and the satisfaction scores.

RTI International, Harvard University, RAND Corporation. *CAHPS Instrument for Persons Residing in Nursing Homes,* Final Report to CMS, CMS Contract No. CMS-01-01176, Sept. 2003.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select Family factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; Family factors should be present at the start of care*)

Not Applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not Applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Analyses used to examine SDS factors include: (1) the summary score for each of the 3 CoreQ: Long-Stay Family questionnaire items; (2) the summary score for the CoreQ: Long-Stay Family measure; and (3) the summary score from the CoreQ: Long-Stay Family questionnaire measure at the facility level.

(1) Summary Score for each of the 3 CoreQ: Long-Stay Family Questionnaire Items

The summary score for each of the 3 CoreQ: Long-Stay Family questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. Correlation and t-test a**nalyses were** used to compare the SDS means with each other. See Table 2b4.4b.a. These analyses show that the individual item scores used in the CORE Q: Long-Stay Family measure are not significantly different based on either education level or race. That is, the educational makeup of the respondents or the racial makeup of the respondents does not appear to relate to the scores for individual items.

Table 2b4.4b.a: Mean CoreQ: Long-Stay Family Distribution Item by Level of Education and Race

What is the highest grade or level of school that you have completed?	Respondents	Q1	Q2	Q3
		Mean	*Mean*	*Mean*
Some high school, but did not graduate	***7% (n=95)***	3.28	3.31	2.50
High school graduate or GED	***32% (n=419)***	3.31	3.45	2.61
Some college or 2 year degree	***32% (n=414)***	3.30	3.44	2.65
4 year college graduate	***15% (n=204)***	3.27	3.42	2.57

More than 4 year college degree	***15% (n=192)***	3.26	3.46	2.61
RANK CORRELATION		0.0056	0.0154	0.0098

RANK CORRELATION OF ITEMS WITH EDUCATION: NONE SIGNIFICANT AT p=0.05

Table 2b4.4b.a: Mean CoreQ: Long-Stay Family Distribution Item by Level of Education and Race (continued)

What is your race?	**Respondents**	**Q1**	**Q2**	**Q3**
		Mean	Mean	Mean
White	***92% (n=1196)***	3.32	3.46	2.63
Black or African-American	**7% (n=92)**	2.98	3.04	2.44
Asian	**1% (n=17)**	3.05	3.47	2.63
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0
TWO-SAMPE T-TEST	1 vs. 2	2.79	3.46	1.59
	1 vs. 3	0.97	0.45	0.49
	2 vs. 3	0.28	1.63	0.77

RACE ITEMS: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(2) Summary Score for the CoreQ: Long-Stay Family Measure

The summary score for each of the 3 CoreQ: Long-Stay Family questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the Family is calculated. Correlation and T-test a**nalyses were** used to compare the SDS means with each other. See Table 2b4.4b.b. These analyses show that the CORE Q: Long-Stay Family measure score is not significantly different based on either education level or race of respondents. That is, the educational makeup of the respondents or the racial makeup of the respondents does not appear to relate to the measure score.

Table 2b4.4b.b: Mean CoreQ: Long-Stay Family Distribution by Level of Education and Race

What is the highest grade or level of school that you have completed?	**Respondents**	**Measure Score**
		Mean
Some high school, but did not graduate	***7% (n=95)***	3.39
High school graduate or GED	***32% (n=419)***	3.66
Some college or 2 year degree	***32% (n=414)***	3.51
4 year college graduate	***15% (n=204)***	3.47
More than 4 year college degree	***15% (n=192)***	3.89

RANK CORRELATION OF MEASURE SCORE WITH EDUCATION: NOT SIGNIFICANT AT p=0.05

Table 2b4.4b.b: Mean CoreQ: Long-Stay Family Distribution by Level of Education and Race (continued)

What is your race?	**Respondents**	**Measure Score**
		Mean
White	***92% (n=1196)***	3.48
Black or African-American	**7% (n=92)**	3.67
Asian	**1% (n=17)**	3.83
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0
American Indian or Alaskan Native	**0% (n=0)**	0
		p-value

TWO-SAMPLE T-TEST	1 vs. 2	0.19
	1 vs. 3	0.21
	2 vs. 3	0.57

RACE MEASURE SCORE: NONE SIGNIFICANTY DIFFERENT AT p=0.05

(4) **Summary score from the CoreQ: Long-Stay Family Measure (at the facility level).**

The summary score for each of the 3 CoreQ: Long-Stay Family questionnaire items is calculated in the following way: Respondents answering poor are given a score of 1, average = 2, good =3, very good =4 and excellent =5. For the 3 questionnaire items the average score for the Family is calculated. The facility score represents the percent of Familys with average scores of 3 or above. A t-test a**nalysis was** used to compare the mean scores. See Table 2b4.4b.c. This analysis demonstrated the CORE Q: Long-Stay Family measure is not significantly different based on either education level or race. That is, the educational makeup of the respondents or the racial makeup of the respondents does not appear to be related to this measure.

Table 2b4.4b.c: CoreQ: Long-Stay Family Score with and without stratification for Education and Race

What is the highest grade or level of school that you have completed?	**Respondents**	**Measure Score**		
		Score with SDS Characteristic vs. Without Characteristic		
Some high school, but did not graduate	***7% (n=95)***	82.5	83.4	n.s
High school graduate or GED	***32% (n=419)***	83.1	83.3	n.s
Some college or 2 year degree	***32% (n=414)***	83.4	82.5	n.s
4 year college graduate	***15% (n=204)***	83.3	83.2	n.s
More than 4 year college degree	***15% (n=192)***	83.9	83.6	n.s

N.S. = Not significant at p=0.05

Table 2b4.4b.c: CoreQ: Long-Stay Family Score with and without stratification for Education and Race (continued)

What is your race?	Respondents	Measure Score (Mean)		
		Score with SDS Characteristic vs. Without Characteristic		
White	***92% (n=1196)***	83.7	83.4	n.s
Black or African-American	**7% (n=92)**	83.5	83.3	n.s
Asian	**1% (n=17)**	83.8	83.5	n.s
Native Hawaiian or other Pacific Islander	**0% (n=0)**	0	0	0
American Indian or Alaskan Native	**0% (n=0)**	0	0	0

N.S. = Not significant at p=0.05

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Not Applicable

Provide the statistical results from testing the approach to controlling for differences in Family characteristics (case mix) below.

If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable

2b4.9. Results of Risk Stratification Analysis:
Not Applicable
2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in Family characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not Applicable

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b).*
We performed an analyses to examine whether the CoreQ Long-Stay Family measure captured clinically/practically meaningful differences between providers by examining a histogram of the scores for the providers in the CoreQ: Long-Stay Family questionnaire sample (Figure 2b5.2.1).
2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
The histogram below shows the distribution of the CoreQ Long-Stay Family measure which has a good distribution and range of scores.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The CoreQ Long-Stay Family scores reflect practical and meaningful differences in quality between facilities. The histogram in Section 2b5.2 shows that the distribution of summary scores is quite wide, indicating the scores can be used to differentiate facilities of varying levels of customer satisfaction quality.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or*

eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not Applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Three items are used in the CoreQ: Long-Stay Family questionnaire. In calculating the CoreQ: Long-Stay Family measure if 1 item of 3 is missing then imputation is used, and if 2 (or more) of the 3 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. The testing to identify the extent and distribution of missing data included examining the frequency of missing responses for each of the 3 CoreQ: Long-Stay Family questionnaire items and the extent and distribution of missing data for more than one missing response for the items. The method of testing to identify if the performance results were biased included examining the correlation with the quality indicators (described above) when imputation was and was not used.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
As noted above, 3 items are used in the CoreQ: Long-Stay Family questionnaire. In calculating the CoreQ: Long-Stay Family measure if 1 item of 3 is missing then imputation is used, and if 2 (or more) of the 3 items is missing, the respondent is excluded. The imputation method consists of using the average score from the items answered. From the testing of 6,192 Family members (described elsewhere) we found:

1. In recommending this facility to your friends and family, how would you rate it overall?
That missing responses occurred in 4.28% (n=265) cases.
2. Overall, how would you rate the staff?
Missing responses occurred in 4.31% (n=267) cases.
3. How would you rate the care your family member received?
Missing responses occurred in 4.25% (n=263) cases.

Two (or more) missing responses occurred in 236 cases. Thus, the degree of missing data was very small (=3.8%). Imputation was used in 220 cases or 3.5% of respondents.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the*

selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data)
Bias from imputation was minimal. The correlation with the quality indicators described above (i.e., restraint use, pressure ulcers, catheter use, antipsychotic use, antidepressant use, antianxiety use, use of hypnotics, and deficiency citations) was unchanged. When the respondents were removed from the analyses, the average Summary Scores remained the same.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Satisfaction Survey

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Since the CoreQ: Long-Stay Family measure has been created and utilized in testing and quality improvement, we have modified it in the following ways.

We conducted analyses on collecting data for the suggested 2 month time period. Even the smallest nursing facilities were able to achieve the 20 survey response goal identified above. We identified that a majority of nursing facilities (i.e., 80%) in our sample could achieve this number of responses if given 2 months. This recommendation was incorporated into the specifications (given above).

As part of the CoreQ: Long-Stay Family measure development, existing satisfaction vendors were contacted (including MyInnerView,

Symbria, and NRC) for input on the administration and sample selection used. With respect to administration, the 2 month window used for including completed surveys are currently often used standard time periods used in the industry. With respect to the sample selection, the exclusion criteria (i.e., residents with court appointed legal guardian for all decisions; residents on hospice) were well received by these vendors. In many cases most of these sample selection criteria are already used by the vendors.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

No fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm) exist.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) AHCA Quality Initiative https://www.ahcancal.org/quality_improvement/qualityinitiative/Pages/Customer-Satisfaction.aspx Satisfaction Vendors N/A Quality Improvement (Internal to the specific organization) Large Nursing Chain N/A

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The measure is currently in use by a large national nursing home chain for the purposes of quality improvement. The data described above was collected from 221 facilities in this chain and included responses from 6,192 family members. These nursing facilities were located in multiple states across the US.

In addition, 10 large national satisfaction vendors in the SNF area have agreed to add the CoreQ to their questionnaires and calculate the measure. The following Customer Satisfaction Vendor are using CoreQ:

•Align
•Brighton Consulting Group
•Healthcare Academy (ReadyQ)
•inQ Experience Surveys
•National Research Corporation (My Innerview)
•Pinnacle
•Providigm/abaqis
•Sensight Surveys

•Service Trac
•The Jackson Group, Inc.

We do not have counts of patients being surveyed and geographical representation from the vendors, however they represent the majority of customer satisfaction vendors currently doing SNF business in the United States.

A letter has been sent to all 10,000 AHCA SNF members indicating which vendors to date have agreed to add the CoreQ to their questionnaire and calculate the measure (see attached letter in appendix, section 4.a.1). A user's manual has been developed and is available on AHCA's website for all satisfaction survey vendors to use. One of the vendors has added the CoreQ to their questionnaire used by States for mandatory satisfaction data collection in all their SNFs (RI, KS and GA), though the results have not yet been calculated by these States.

AHCA and NCAL have also incorporated the CoreQ into their national Quality Initiative goals. AHCA represents nearly 10,000 of the 15,000 SNFs and provides feedback to all of its members on their satisfaction scores using the CoreQ. This has resulted in growing number of members and vendors collecting the data.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The CoreQ: Long-Stay Family questionnaire measure is not currently publicly reported or used in other accountability applications (e.g., payment program, certification, licensing). The reason for this is that it is a new measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
AHCA has recently started the second Quality Initiative, laying out a series of quality improvement and reporting goals for the AHCA membership, which covers nearly 10,000 of all 15,000 SNFs in the U.S. Among these goals is the collection and reporting of CoreQ customer satisfaction data. Because it has been included in the Quality Initiative 2015-2018, AHCA's machinery for publicizing and encouraging the adoption of the tool has been activated, including AHCA's quality division spending a large number of staff hours working to accomplish this. In addition to marketing the use of the survey instrument as a way for SNFs to understand how their patients view the care and other services that they were provided by the SNFs, AHCA is developing an upload and reporting feature within its member data profiling tool, LTC Trend TrackerSM, which allows SNFs to centrally view a large number of quality, compliance, operational and financial metrics from public and non-public sources. The CoreQ report and upload feature within LTC Trend Tracker will include an API for vendors performing the survey on behalf of SNFs – AHCA's preferred approach to collecting the data – so that the aggregate CoreQ results will be immediately available to providers as they are collected. Given that LTC Trend TrackerSM is the leading method for SNFs to profile their quality and other data, the incorporation of CoreQ into LTC Trend Tracker means it will immediately become the de facto standard for customer satisfaction surveys for the SNF industry.

We also are working with states who require satisfaction measurement to incorporate the CoreQ into their process. The State of Rhode Island and Georgia pilot tested a version of the CoreQ in its statewide satisfaction questionnaire for long stay residents. The vendor in KS will also be adding the CoreQ to their questionnaire. The state of Massachusetts has included the CoreQ short stay as part of its current ongoing deliberation on measuring satisfaction in SNFs. AHCA has a presence in each state, and our state affiliates will be promoting the use of the CoreQ in those states that are collecting or considering collecting satisfaction.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not Applicable.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of**

initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
Not Applicable.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no negative consequences to individuals or populations identified during testing or evidence of unintended negative consequences to individuals or populations reported since the implementation of the CoreQ: Long-Stay Family questionnaire or the measure that is calculated using this questionnaire. This is consistent with satisfaction surveys in general in nursing facilities. Many other satisfaction surveys are used in nursing facilities with no reported unintended consequences to patients or their families.

There are no potentially serious physical, psychological, social, legal, or other risks for patients. However, in some cases the satisfaction questionnaire can highlight poor care for some dissatisfied patients, and this may make them further dissatisfied.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0693 : Consumer Assessment of Health Providers and Systems (CAHPS®) Nursing Home Survey: Family Member Instrument

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
N/A

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The CoreQ: Long-Stay Family measure does not conceptually address the same measure focus as any other NQF-endorsed measures, however it does conceptually address the same target population as another NQF-endorsed satisfaction measure. The Consumer Assessment of Health Providers and Systems (CAHPS®) Nursing Home Family Member Survey Instrument (NQF #0693) presented by the Agency for Healthcare Research and Quality received NQF approval over five years ago in March, 2011. This instrument is endorsed to collect family member satisfaction information and consists of a 50 item questionnaire. Our application also uses nursing home residents (The CoreQ: Long-Stay Family measure) but consists of three items that are aggregated into a single measure. The score from these items is used to provide standardized information on the overall family satisfaction of the facility. The current CAHPS survey is not used in this way.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** CoreQ_Family_Appendix_Final.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Health Care Association
Co.2 **Point of Contact:** Urvi, Patel, upatel@ahca.org, 202-898-2858-
Co.3 **Measure Developer if different from Measure Steward:** American Health Care Association
Co.4 **Point of Contact:** Lindsay, Schwartz, lschwartz@ncal.org, 202-898-2848-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The workgroup gave input, reviewing our suggested administration, required response rate, the manual, and exclusions.

Mary Tess Crotty, Genesis - Also helped provide feedback on the development process and the user manual. Additionally, she reviewed the analyses.
Matt O'Connor HCR Manor Care- Also helped provide feedback on the development process and the user manual. Additionally, he conducted some analyses and reviewed the analyses.
Judy Hoff, Health Care Academy
Rich Kortum, My Innerview/National Research Corporation
Peter Kramer, abaqis/Providigm
Ellen Kuebrich, abaqis/Providigm
Michael Johnson, ServiceTrac
Chris Magelby, Pinnacle

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.6 **Copyright statement:** None
Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2769
Measure Title: Functional Change: Change in Self Care Score for Skilled Nursing Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Brief Description of Measure: Change in rasch derived values of self-care function from admission to discharge among adult patients treated as short term rehabilitation patients in a skilled nursing facility who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
Developer Rationale: The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The self-care measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. While the self-care measure is not required as part the MDS system used in SNFs, currently more than 150 SNFs are collecting data on the items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

Numerator Statement: Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
Denominator Statement: Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age
Denominator Exclusions: Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. They provide a list of 9 peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in SNFs.
- In addition, they provide summaries/abstracts from three articles that support the following: *The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay.*
- The items in the self-care score are: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

According to the developer, "The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers."

This is a new measure, but UDSMR has been collecting data on the FIM instrument for 20 years, so they are able to report on trends. Almost half (46%) of facilities were below expectation in 2014:

Year	2010	2011	2012	2013	2014
Selfcare Change Average (Rasch)	17.6	17.4	17.0	16.7	16.6
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation (1.0)	62	66	76	67	83
Number of Facilities below Expectation (< 1.0)	66	75	71	76	71
Percent of Facilities at or above Expectation (1.0)	48.4%	46.8%	51.7%	46.9%	53.9%

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the motor

measure for the years 2010 to 2014. Information is not provided on whether the differences are statistically significant, however, the data to provide information on factors for consideration in assessing variation and impact on various populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**The evidence provided appears to be for the FIM - which is the tool used to provide the items that make up the measure.

**This is an important area for measurement as the population ages and more patients may be admitted to SNF's for rehab.
Data was supplied for evidence of the FIM tool for successful rehab. The components of the tool haven't been studied separately for outcomes specific to the self-care components being specified in this measure.
I was also wondering about the timeframe for the measure being 12 months. Most short term stays are much shorter, weeks to a few months and I would be interested in knowing how the 12 month timeframe was decided.

**There is a clear link between a SNF's performance of services and the outcome measured at the patient level.
The journal articles demonstrate that the primary goal of rehabilitation is to restore or improve function, but Jimmo ensures that skilled nursing and therapy must be provided to patients to maintain and/or prevent deterioration of function as well. How does this measure accommodate for this requirement?
The measure results involve ADLs that can clearly be impacted by SNF services, mainly through the provision of skilled therapy/nursing as well as patient education.
Agree with the pass on evidence but the Jimmo omission is problematic.

1b. Performance Gap

Comments:

**I wasn't clear what the performance gap actually was - the differences in FIM that might indicate improved performance are actually quite small – it's not clear how meaningful they would be, and differences are very dependent on the individuals admitted to the SNF in the first place. However, there does appear to be some variability in the measure.
**There was data provided that looked at variation and opportunity. About 50% of the facilities are at their expected performance.

SDS data was provided but no interpretation as to whether there is any statistical differences.

**This is a new measure but UDS has longitudinal data that allows them to analyze the effectiveness of this measure in demonstrating the difference in expected outcomes vs. achieved outcomes over previous years in SNFs.

The fact that almost half (46%) of SNFs were below expectations in 2014 shows a significant gap in care that could benefit from such an outcome measure at a national level.

The disparities data could be more detailed and comprehensive but UDS did make efforts to provide some data on disparities in care based on gender, race, payer source, and region.

I agree with the staff recommendation for a moderate rating.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, MDS data, and SNF CMG codes (case mix group)

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:
  - Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients, over Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.
  - Average is calculated as (sum of change at the patient level/total number of patients).
- The calculation algorithm is included.
- Patients under age 18 and patients who died in the SNF are excluded.
- A data dictionary is included.
- The measure is stratified by risk category.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented
- This measure uses a subset of the FIM, so a Rasch analysis was conducted to test:
    - the psychometric properties of the subset of 8 items within the three venues of post-acute care, IRFs, LTACs, and SNFs
    - The measure reliability at both the person and item level
    - to determine the fit of each item within the measure (8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) through infit and outfit statistics and item specific correlations.
- Internal consistency of the critical data elements was demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but us striving to do so prior to the in—person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities (perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities.* UDSMR has indicated they are working to fulfill these data needs.

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.70 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empirical testing of data elements (box 2) -> TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity Testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Developers used concurrent validity of the FIM total score (all 18 items) with the FIM self-care score. Specifically the following analyses were conducted: the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (self-care items).
- Predictive validity of the self-care score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:

- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for SNFs are 0.937 ($p < 0.001$) at admission and 0.871 ($p < 0.001$) at discharge.
- For predicative validity of functional gain, SNFs scored 0.681 ($p < 0.001$) and an r-squared value of 0.464.
- For SNFs, the r-squared values at admission were 0.877 and at discharge 0.758. The C-statistic for SNFs is 0.80.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the facility were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developer states the following risk adjustment method: To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same SNF-CMG are expected to have similar resource utilization needs and similar outcomes.

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Skilled Nursing Facility Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *No conceptual basis for adjusting this measure for SDS factors is included. Do you think it should be?*
- *No information is provided on risk adjustment for SDS factors. Do you think the measure should include SDS factors in the risk adjustment? Why or why not?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provided additional information in an addendum, including "graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers".

Question for the Committee:

Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 Specifications

Comments:

**The way the measure is calculated from the FIM is explained well. However, I have concerns about the implementation - the FIM is not routinely used in SNFs. A number of the elements it uses are similar to the measures collected currently in the MDS and the CARE tool - so I would be worried about data burden if we expect SNFs to collect this data as well as the CMS mandated data.

**The specifications given are components of the overall FIM tool.
Analysis was done to see if the self-care components correlate with the FIM tool and there was correlation.
One of the studies referenced did look at ADL's and mobility and concluded the patterns of functional change differed for these domains and for specific groups of patient - this was at the patient level, not facility level (Latham et al).
Are there different components of the FIM tool that are more predictive of improvement/discharge potential?

Another consideration is that staff need to be trained to use the tool and is there consistency in the training and evaluation as to the skills of the staff in using this tool?

**Data on reliability is incomplete and UDS is expected to release new data analysis on reliability by the time of the meeting. However, the longitudinal data and the eight elements being measured are well established and reliable indicators of functional improvement (The Cronbach Alpha score was .92, which is very high).

The data thus far does suggest that the test sample is adequate to generalize for widespread implementation.

2a.2 Validity Testing

Comments:

**I wasn't sure what the ICC was telling me with regard to the reliability of the measure (as opposed to the items that make up the measure). My understanding/interpretation of their result would be that the measure doesn't have consistency across facilities - although they seem to be saying that actually it means that there is inconsistency and that is good????
**Internal consistency for the facilities was demonstrated.
The assessment of reliability across facilities showed the ICC of a negative value with a high p value - it might be helpful to have further explanation of those results.

2b.2 Validity Testing

Comments:

**The measure does appear to have predictive validity.
**Facility testing was done. A number of facilities were included.
An observation is that the number of facilities tested increased from 2010 to 2014 from 128 to 154 while the case count decreased from 26472 to 21629. It would help to know the reason for this as it would seem if the number of facilities increased the case count would as well.
**Method of validity testing of the measure score was empirical validity testing, not face validity.

Predictive validity of the self-care score was tested to ensure a connection between the measure and the outcome to be achieved (i.e., functional change and likelihood of discharge to the community)

Question: Is this measure to be applied across IRFs, LTACHs, and SNFs, or just SNFs? It appears to have been tested in all three settings but the application confines the measure's use to SNFs only.

The validity testing that was conducted does indicate that conclusions about quality can be made using this measure but with one reservation. I believe this measure is a valid measure of functional outcome for those who achieve improvement during the course of their SNF stay. But if patients do not improve, I have concerns that it will penalize SNFs for accepting patients who need skilled therapy to maintain or prevent deterioration of function. What can be done to mitigate against this bias?

2b.3-7. Exclusion Analysis

Comments:

**I was slightly confused by the risk adjustment - the measure itself is for patients admitted to SNFs for short term rehabilitation, but the risk adjustment appears to be on the whole SNF population - and I think I'd like more reassurance that this measure is appropriate to risk adjust for patients targeted by the measure.

**Exclusions include children and patients that died.

A comment is made that patients who don't have complete data are deleted from the measure. The impact to the measure isn't clear - does this happen in only a few facilities, in only a few patients etc.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- The Functional Change: Change in Motor Score form (this form includes the items for the self-care measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**I have concerns regarding duplication of data collection, given the overlap in data elements between the FIM (which is used to collect the data used for the measure) and CMS mandated data that is already collected by CMS. In reality the number of SNFs that use the FIM at present is very small, and there is a potential cost to facilities for using it.

**The tool and measurement are part of a subscription to UDSMR.

It isn't clear if this measure could be done outside of that structure if a facility wanted to purchase only the FIM tool.

**The required data elements are already collected and this measure relies heavily on electronic submission of data.

The UDS measure is proprietary so this may undercut the widespread use of the measure across the nation. For this reason, I agree with the moderate rating. If the measure were not proprietary, I would recommend high feasibility.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

- The measure is currently used for internal reporting and national benchmarking by SNFs who subscribe to the UDSMR software/outcomes reporting.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results:

- New measure – not available. While a new measure to NQF, the developer does provide trending data for the rasch derived scores back to 2010:

Year	2010	2011	2012	2013	2014
Selfcare Change Average (Rasch)	17.6	17.4	17.0	16.7	16.6
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation (1.0)	62	66	76	67	83
Number of Facilities below Expectation (< 1.0)	66	75	71	76	71
Percent of Facilities at or above Expectation (1.0)	48.4%	46.8%	51.7%	46.9%	53.9%

Unexpected findings (positive or negative) during implementation
None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4 Usability

Comments:

**The measure is being used for internal quality improvement and national benchmarking across the SNFs that use the FIM currently.

**The measure is being proposed for use in accountability programs. An unintended consequence could be longer length of stays to allow patients to have a higher score. This could lead to increased costs. A measure of improvement would help patients/facilities and could be even stronger if there was a balancing metric on length of stay or costs of treatment.

**The benefits of the measure outweigh any unintended consequences because the data is already being collected (or could be easily collected by SNFs) electronically. The measure is a critical adjunct to existing measures in the SNF setting because one of the core reasons for SNF care is functional status, which is largely not measured currently under existing measurement tools.

The gap between functional gains and expected functional gains will enable SNFs to better self-assess their own performance in the primary area they are expected to measure, functional gain and, ultimately, ability to return to the community after a short SNF stay. In this manner, this measure will enable benchmarking and clearly furthers the goal of high-quality, efficient and EFFECTIVE health care.

Criterion 5: Related and Competing Measures

Related or competing measures

2613 : CARE: Improvement in Self Care

Harmonization

None

Pre-meeting public and member comments

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Self Care Score for Skilled Nursing Facilities

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

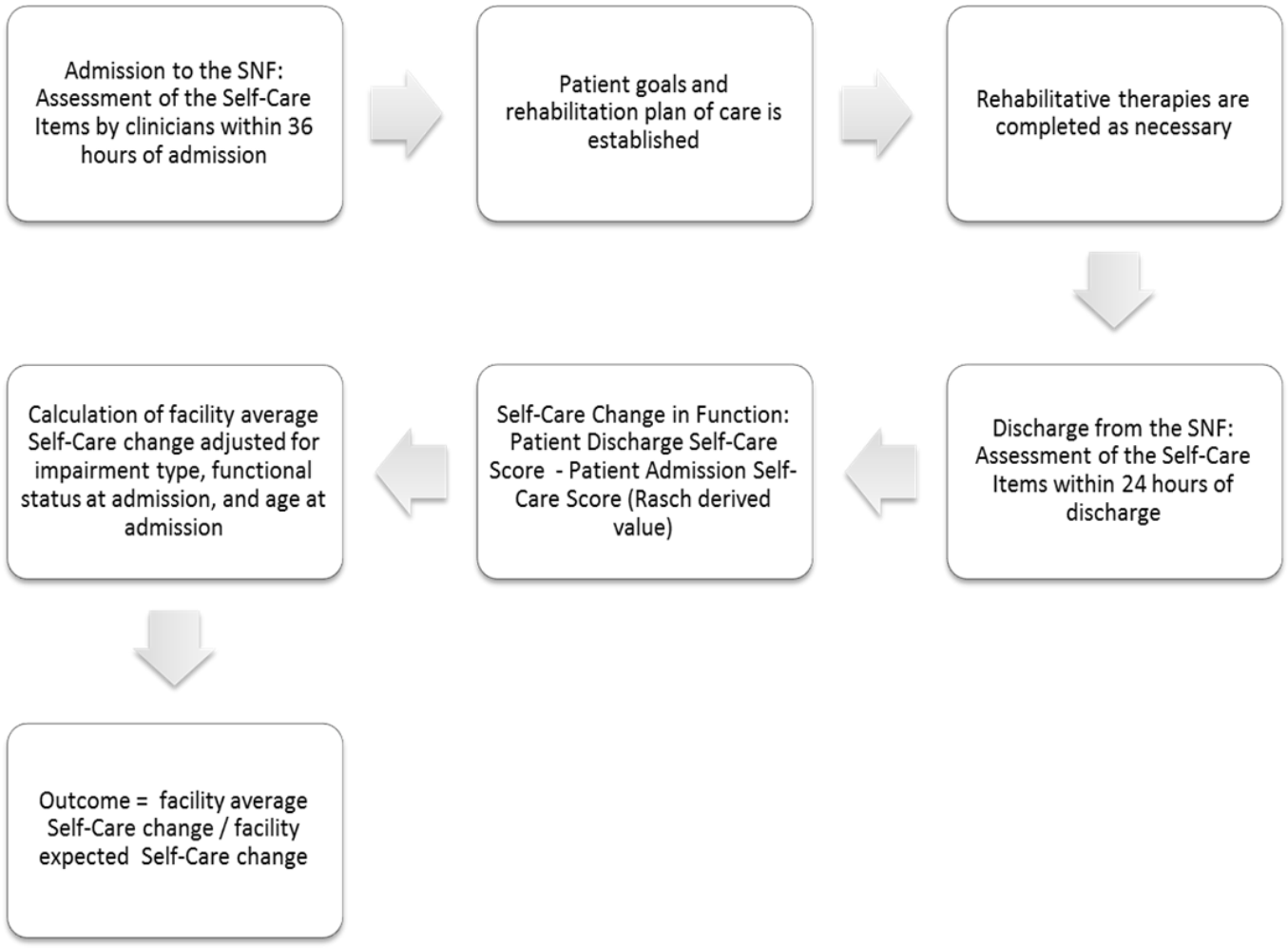
☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Skilled Nursing Facilities (SNFs) are one part of a multi-level post-acute care continuum. Two different types of patients are admitted to SNFs; those meant to live in the facility, and those to receive short-term rehabilitation. The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are over 150 SNFs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for short term rehabilitation patients seen at SNFs. To date, the self-care measure has not been reported on as a stand-alone measure. However, the items of the self-care measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument.. The self-care measure is intended to be administered within 24 hours of the patient's admission to the IRF and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the

self-care measure are as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory). All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for SNF short term rehabilitation patients:



While the self-care measure is new, UDSMR has been a data repository for the FIM instrument among SNF patients, of which the items of the self-care measure are nested within for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the self-care measure for the years 2010 to 2014 for short term skilled nursing facility patients:

Year	2010	2011	2012	2013	2014
Selfcare Change Average (Rasch)	17.6	17.4	17.0	16.7	16.6
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation (1.0)	62	66	76	67	83
Number of Facilities below Expectation (< 1.0)	66	75	71	76	71
Percent of Facilities at or above Expectation (1.0)	48.4%	46.8%	51.7%	46.9%	53.9%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the mobility measure for the years 2010 to 2014.

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2010		2011		2012		2013		2014	
	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)
Gender										
Male	7,668	17.5	7,705	17.2	7,617	17.0	6,489	16.6	5,100	16.8
Female	13,768	17.9	13,730	17.6	13,061	17.2	10,362	17.1	8,204	16.9
Ethnicity										
White	14,461	17.5	14,422	17.1	13,586	17.0	9,766	16.8	8,014	16.5
Black	2,073	16.3	2,273	17.9	1,997	17.5	1,609	17.0	1,453	16.7
Hispanic	370	19.0	400	17.3	353	17.8	216	16.8	140	16.3
Other Ethnicity	9,568	17.8	9,559	17.5	10,991	17.0	14,029	16.7	12,022	16.7
Payer Source										
Medicare	18,658	17.5	19,261	17.3	19,898	16.9	18,842	16.6	15,577	16.5
Medicaid	669	14.3	525	17.1	566	18.1	519	17.2	514	17.8
Commercial	1,826	17.3	2,032	17.6	2,052	17.0	2,247	16.9	1,799	16.6
Blue Cross	1,168	21.3	845	20.9	876	20.0	999	18.7	526	18.2
Other Payer	4,151	17.3	3,991	16.9	3,535	16.9	3,013	16.7	3,213	16.5
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	3,481	17.5	3,310	16.6	3,784	16.6	3,539	16.5	3,437	16.1
P02 (NY, NJ, PR)	9,099	19.5	7,581	19.0	6,031	18.8	6,290	17.9	4,426	17.4
P03 (PA, WV, VA, DE, MD, DC)	1,793	16.5	1,489	16.7	1,565	18.5	1,721	16.7	1,198	16.1
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	8,057	15.7	7,542	16.9	7,401	16.0	8,759	15.6	7,405	15.8
P05 (MN, WI, IL, IN, MI, OH)	3,728	17.7	3,290	17.0	3,313	17.7	4,289	17.7	4,907	17.7
P06 (NM, OK, AR, LA, TX)	29	15.6	2,015	16.0	2,685	15.4	383	15.2	0	-
P07 (NE, IA, KS, MO)	285	16.2	1,381	15.6	2,124	16.3	639	16.1	135	14.8
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	33	16.5
P09 (CA, NV, AZ, HI)	0	-	46	21.4	24	18.3	0	-	88	17.9
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the self-care measure is a new measure and has not been used as a stand-alone tool. However, all of the items within the measure are included in a larger instrument, the FIM instrument, which has

been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the self-care measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and use of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Skilled Nursing Facilities:

1. Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.

2. Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.

3. Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.

4. Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

5. Munin MC, Begley A, Skidmore ER, Lenze EJ. Influence of rehabilitation site on hip fracture recovery in community-dwelling subjects at 6-month follow-up. *Archives of physical medicine and rehabilitation.* Jul 2006;87(7):1004-1006.

6. Munin MC, Seligman K, Dew MA, et al. Effect of rehabilitation site on functional recovery after hip fracture. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):367-372.

7. Nelson DL, Melville LL, Wilkerson JD, Magness RA, Grech JL, Rosenberg JA. Interrater reliability, concurrent validity, responsiveness, and predictive validity of the Melville-Nelson Self-Care Assessment. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association.* Jan-Feb 2002;56(1):51-59.

8. Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Archives of physical medicine and rehabilitation.* Oct 1996;77(10):1056-1061.

9. Vincent KR, Vincent HK. A multicenter examination of the Center for Medicare Services eligibility criteria in total-joint arthroplasty. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Jul 2008;87(7):573-584.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.

OBJECTIVE: To examine the outcomes of patients with varying levels of cognitive impairment who received rehabilitation in skilled nursing facilities (SNFs). DESIGN: A retrospective analysis of the records of people admitted to SNFs for rehabilitation. SETTING: Seven SNFs in Colorado. PARTICIPANTS: Community-dwelling persons (N=7159), 65 years of age and older, admitted for rehabilitation after a hospitalization or decline in function between May 1998 and May 2002. Interventions Not applicable. MAIN OUTCOME MEASURES: Cognitive impairment was assessed using a 4-level categorization of the FIM instrument cognitive score at admission. Functional gain was measured using the FIM. Community discharge was measured as the proportion of patients discharged to home, board and care, or assisted living facility. Rehabilitation progress was measured as the number of FIM points gained per day. RESULTS: Significant functional gains were made during rehabilitation in motor and cognitive FIM scores, regardless of cognitive impairment. The most cognitively impaired patients required more rehabilitation intervention, achieved less FIM gain, and were less likely to be discharged to the community. The strongest predictors of FIM gain were the amount of therapy hours and admission cognitive FIM score. The strongest predictors of discharge to the community were the discharge total FIM score and age. The strongest predictors of adequate rehabilitation progress were medical complexity and admission cognitive FIM score. CONCLUSIONS: Patients with cognitive impairment were able to recover function with rehabilitation intervention. Patients with a more serious cognitive impairment received more rehabilitation intervention than patients with less impairment. Outcomes were predicted by admission and rehabilitation measures that were qualitatively different from other discharge outcomes. Health care professionals need to consider these factors as they create a rehabilitation plan of care for patients with cognitive impairment.

Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.

OBJECTIVES: To document patient, program characteristics, and therapy service provision in subacute rehabilitation across 3 types of facilities that provide subacute rehabilitation, to examine the determinants of therapy intensity, and to evaluate the contribution of rehabilitation services to functional gains. DESIGN: A retrospective study linking administrative billing data and patients' functional assessment records. SETTING: Twenty facilities part of the Uniform Data System for Medical Rehabilitation (UDSMR) subacute database PARTICIPANTS: A total of 1976 billing records of patients with stroke, orthopedic, and debility impairments, discharged in 1996 and 1997, were retrieved and linked with the FIM trade mark instrument ratings from UDSMR subacute database. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: Total therapy intensity and Rasch-transformed FIM domain gains (ie, gains in self-care, mobility, cognition). RESULTS: Therapy intensity was mostly determined by impairment and facility type, although variances explained by the predictors were small. Patients in all 3 impairment groups made functional gains; gains were related weakly, although significantly, to therapy intensity and rehabilitation duration after controlling for other variables. CONCLUSIONS: The provision of rehabilitation therapies varied across facilities. Skilled nursing facilities with subacute rehabilitation units tended to provide more therapies than subacute units in acute or rehabilitation hospitals.

Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.

OBJECTIVE: To examine the relation between therapy intensity, including physical therapy (PT), occupational therapy (OT), and speech and language therapy (SLT), provided in a skilled nursing facility (SNF) setting and patients' outcomes as measured by length of stay (LOS) and stage of functional independence as measured by the FIM instrument. DESIGN: A retrospective analysis of secondary data from an administrative dataset compiled and owned by SeniorMetrix Inc. SETTING: Seventy SNFs under contract with SeniorMetrix health plan clients. PARTICIPANTS: Patients with stroke, orthopedic conditions, and cardiovascular and pulmonary

conditions (N=4988) covered by Medicare+Choice plans, and admitted to an SNF in 2002. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: LOS and improvement in stage of independence in the mobility, activities of daily living (ADLs), and executive control domains of function as determined by the FIM instrument. RESULTS: Higher therapy intensity was associated with shorter LOS (P <.05). Higher PT and OT intensities were associated with greater odds of improving by at least 1 stage in mobility and ADL functional independence across each condition (P <.05). The OT intensity was associated with an improved executive control stage for patients with stroke, and PT and OT intensities were associated with improved executive control stage for patients with cardiovascular and pulmonary conditions (P <.05). The SLT intensity was associated with improved motor and executive control functional stages for patients with stroke (P <.05). Therapy intensities accounted for small proportions of model variances in all outcomes. CONCLUSIONS: Higher therapy intensity was associated with better outcomes as they relate to LOS and functional improvement for patients who have stroke, orthopedic conditions, and cardiovascular and pulmonary conditions and are receiving rehabilitation in the SNF setting.

Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

OBJECTIVE: To examine the rate of functional change in 2 domains, activities of daily living (ADLs) and mobility, over 2 time periods during hip fracture rehabilitation. DESIGN: Retrospective analysis of data contained in an administrative dataset. SETTING: Seventy skilled nursing facilities (SNFs). PARTICIPANTS: People (N=351) receiving rehabilitation in SNFs from March 1998 to February 2003 after hip fractures. INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURE: Rate of change in scores in the ADL and mobility domains of the FIM instrument during 2 time intervals of rehabilitation. RESULTS: The rate of functional change across 2 time intervals was constant for mobility (mean change in FIM points per day, .46 vs .49), but declined in the second time period for ADLs (mean change in FIM points per day, .55 vs .41). Executive function, length of stay (LOS), and medical complexity were related to rate of change in mobility, and baseline ADLs, executive function, living setting, and LOS were related to rate of change in ADLs. There was an interaction between rehabilitation phase and baseline mobility. People with lower baseline mobility had an increased rate of change during the second interval (mean change in FIM points per day, .41 vs .55), whereas those with higher baseline mobility had a decreased rate of change (mean change in FIM points per day, .50 vs .43). CONCLUSIONS: The pattern of functional change over time differed for ADL and mobility domains, and for specific groups of patients. The results have implications for goal setting and discharge planning.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2769
De.2. **Measure Title:** Functional Change: Change in Self Care Score for Skilled Nursing Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
De.3. **Brief Description of Measure:** Change in rasch derived values of self-care function from admission to discharge among adult patients treated as short term rehabilitation patients in a skilled nursing facility who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
1b.1. **Developer Rationale:** The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The self-care measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. While the self-care measure is not required as part the MDS system used in SNFs, currently more than 150 SNFs are collecting data on the items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

S.4. **Numerator Statement:** Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
S.7. **Denominator Statement:** Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age
S.10. **Denominator Exclusions:** Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Self_Care_SNF-635950326281534154.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The self-care measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. While the self-care measure is not required as part the MDS system used in SNFs, currently more than 150 SNFs are collecting data on the items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.

We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Please see Measure Evaluation Form for data over time

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Please see Measure Evaluation Form for disparities data

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
In the most recent MedPAC report (June 2015), there were over 2 million stays in SNFs among Medicare beneficiaries alone. In addition, it has been noted at the government level that function is imperative when discussing quality of care among post-acute care venues.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/IMPACT-Act-of-2014-and-Cross-Setting-Measures.html

MedPAC. Health Care Spending and the Medicare Program June 2015: http://medpac.gov/documents/data-book/june-2015-databook-health-care-spending-and-the-medicare-program.pdf?sfvrsn=0

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular, Cardiovascular : Ischemic Heart Disease, Coronary Artery Disease, Musculoskeletal, Musculoskeletal : Hip/Pelvic Fracture, Musculoskeletal : Joint Surgery, Musculoskeletal : Low Back Pain, Musculoskeletal : Osteoarthritis, Musculoskeletal : Osteoporosis, Musculoskeletal : Rheumatoid Arthritis, Neurology, Neurology : Brain Injury, Neurology : Cognitive Impairment/Dementia, Neurology : Stroke/Transient Ischemic Attack (TIA)

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_Submission_Self_Care_SNF.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 Months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in the SNF. The numerator is the average change in rasch derived self-care functional score from admission to discharge for each patient at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory. Average is calculated as: (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in the SNF. Impairment type is defined as the primary medical reason for the SNF short term rehabilitation stay (such as stroke, joint replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum of 8 items ((Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) at the facility level. Age is the age of the patient at the time of admission to the SNF. The denominator is meant to reflect the expected Self-Care functional change score at the facility, if the facility had the same distribution of SNF-CMGs (based on impairment type, functional status at admission, and age at admission). This adjustment procedure is an indirect standardization procedure (observed facility average/expected facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through the MDS.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the SNF-CMGs in the appendix.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standarization procedure (observed facility average/expected facility average). The numerator is the facility´s average self-care functional change score. The denominator is meant to reflect the expected Self-Care functional change score at the facility, if the facility had the same distribution of SNF-CMGs (impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all short term rehabilitation patients during the assessment time frame (12 months).
2. Exclude any patients who died in the SNF.
3. Exclude any patients who are less than 18 at the time of admission to the SNF.
3. Calculate the total self-care change score for each of the remaining patients (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in attached excel file).
5. Calculate the average rasch derived self-care change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average self-care change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average self-care change score/facility´s national expected self-care change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the appendix.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Self_Care_SNF-635950326454202907.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Self Care Score for Skilled Nursing Facilities Click here to enter measure title
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the

information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors

between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® ("FIM") instrument data from inpatient rehabilitation facilities, long term acute care facilities, and skilled nursing facilities from the Uniform Data System for Medical Rehabilitation. The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the self-care measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2010 - 2014 were used in examining the data trends over time using the self-care measure and patient outcomes of skilled nursing facilities

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: patient level, aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term acute care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument (the tool used for this measure) is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM® instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 8 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for the inpatient rehabilitation setting. However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our self-care measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistant compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.70 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure the Rasch scale was set to 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.

```
EXPECTED SCORE: MEAN  (Rasch-score-point threshold, ":" indicates Rasch-half-point
threshold) (ILLUSTRATED BY AN OBSERVED CATEGORY)
0    10    20    30    40    50    60    70    80    90
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   Item
1                 1  :   2 :3 : 4:  5 :   6    :    77    4  DressingLower
1                1  :   2 :3 : 4:  5 :   6    :    7 7    5  Toileting
|                                                      |
|                                                      |
1             1   :  2 : 3 :4 : 5  :   6    :   7      7    3  DressingUpper
1           1   :  2 : 3: 4 : 5  :  6    :    7        7    6  Bowel
1          1   :  2 : 3: 4 : 5  :  6    :    7         7    2  Grooming
|                                                      |
1        1   :  2 : 3: 4 : 5  :  6    :    7           7    8  Memory
1       1  :   2 :3 :4 : 5  :   6    :   7             7    1  Eating
|                                                      |
1     1   :  2 : 3: 4 : 5  :  6    :    7              7    7  Expression
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   Item
0    10    20    30    40    50    60    70    80    90
```

The ruler shows that the easiest functional item is Expression, and the most challenging functional item is Dressing Lower, additionally, the distances between a level 1 and 2 and 5, 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 8 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
8	.00E	19.37	25	38.91	3.03	42	54.11	3.50
9	11.92	10.11	26	39.75	3.01	43	55.26	3.61
10	18.17	6.88	27	40.58	3.00	44	56.50	3.72
11	21.64	5.55	28	41.40	3.00	45	57.81	3.86
12	24.08	4.81	29	42.22	3.00	46	59.23	4.01
13	25.99	4.35	30	43.05	3.00	47	60.77	4.18
14	27.60	4.03	31	43.88	3.01	48	62.44	4.37
15	29.00	3.80	32	44.71	3.02	49	64.28	4.60
16	30.26	3.62	33	45.56	3.04	50	66.33	4.88
17	31.42	3.49	34	46.41	3.06	51	68.67	5.24
18	32.50	3.38	35	47.28	3.09	52	71.42	5.73
19	33.52	3.29	36	48.17	3.12	53	74.81	6.48
20	34.49	3.22	37	49.07	3.16	54	79.39	7.78
21	35.42	3.16	38	50.00	3.21	55	86.98	10.87
22	36.33	3.12	39	50.97	3.27	56	100.00E	19.83
23	37.20	3.08	40	51.97	3.34			
24	38.07	3.05	41	53.01	3.41			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*).

The results of the analysis for the self-care measure were statistically significant, the Cronbach's alpha indicated very high internal consistency, thus a very stable measure.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the self-care measure with the FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new self-care measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM® total) explained by the independent variable (self-care items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the self-care measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM® instrument (the gold standard)), and likelihood of discharge to the community setting Linear regression was used to determine functional change, whereas the change in self-care was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission self-care total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no)t. We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients s having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the self-care measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was highly correlated with the FIM total, 0.929 ($p < 0.001$) and 0.881 ($p < 0.001$), respectively. The correlations remained significant within each venue of care; IRFs, 0.933 ($p < 0.001$) and 0.896 ($p < 0.001$); LTACs, 0.928 ($p < 0.001$) and 0.888 ($p < 0.001$); SNFs, 0.937 ($p < 0.001$) and 0.871 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM totals, the r-square values were very high for admission FIM total and discharge FIM total, 0.864 and 0.775, respectively. The values remained similar at the venue specific level as well; IRFs, 0.870 and 0.804; LTACs, 0.861 and 0.788; SNFs, 0.877 and 0.758.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was strong, 0.721 ($p < 0.001$). In addition, by venue, the correlations remained strong; IRFs, 0.780 ($p < 0.001$); LTACs, 0.757 ($p < 0.001$); SNFs, 0.681 ($p < 0.001$). The linear regression showed significant, high r-squared values as well; all venues, 0.519; IRFs, 0.608; LTACs, 0.574; SNFs, 0.464.
Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in self-care has good predictive ability for discharge setting (community), with a C-statistic of 0.76. By venue, the results are similar; IRFs, 0.74; LTACs, 0.73; SNFs, 0.80.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results show the self-care measure is valid; the measure demonstrated construct, concurrent, discriminant and predictive validity in all analyses. The r-square values were all consistent, 0.6 or higher, meaning that the percent of variance explained in the dependent variables by our measure were all more than 60%. The predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients that died in the post-acute care setting (an unanticipated outcome) and patient aged 18 years and older, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
No statistical tests completed.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 1 risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. SNF-CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM instrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all SNF-CMGs.)

See file uploaded in S.15 for calculations.

The SNF-CMGs are groupings specific to skilled nursing facilities, although they are similar and easily comparable to the CMGs used in inpatient rehabilitation facilities.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, SNF-CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a*

method; what statistical analysis was used)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.
2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.
4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Over 150 SNFs currently use UDSMR and the FIM instrument for quality benchmarking, both internally and as a national benchmarking system. The self-care measure is embedded in the full FIM instrument. Therefore, the feasibility of this measure is sound.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

The Functional Change: Change in Motor Score form (this form includes the items for the self-care measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a

license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Uniform data System for Medical Rehabilitation www.udsmr.org Quality Improvement (Internal to the specific organization) Uniform data System for Medical Rehabilitation www.udsmr.org

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for SNFs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for skilled nursing facilities, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

The items of our proposed measure are part of the FIM system, which is in use in nearly 150 SNFs in the United States. Outcomes based on the items are currently used for Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and Quality Improvement (Internal to the specific organization) for those SNFs utilizing the FIM system.

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for*

implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
There were no unintended negative consequences to individuals or populations during the testing of this measure as previously collected data was used.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)
2613 : CARE: Improvement in Self Care

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
While the CARE items and the self-care measure the same construct of functional (in)dependence, there are some key differences key differences included in the measures, and in the measurement of the items. The self-care measure submitted by UDS includes the following items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory. The CARE items included in the measure submitted by AHCA include: Eating, Oral hygiene, Toilet hygiene, Shower/bathe self, Upper body dressing, Lower body dressing, Putting on/taking off footwear. Once again there is great overlap in the items, particularly for feeding, grooming, and toileting. However, where the AHCA measure does not contain any cognitive items in their measure, our measure contains two cognitive items when determining a patient's ability to care for one's self especially for discharge planning, cognitive ability play a key role, thus we maintain our measure is best in class considering it is more robust, has greater sensitivity in measurement (our measure uses a seven level rating scale whereas the CARE measure uses a six level, thus our rating scale offers greater refinement in measurement). Finally, the UDSMS change in self-care measure is the exact same measure (same items, same rating scale, same adjustment) used in SNF, IRF and LTAC, offering consistency in measuring patient function across PAC venues, which has been an interest for PAC and is a current objective of the IMPACT ACT.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
The functional items in our proposed measure have been collected in SNFs for over 20 years. This allows for a historical perspective of function in the SNFs that the CARE items do not allow. In addition, the functional items in our proposed measure have been used in inpatient rehabilitation facilities for over 30 years, and therefore, a comparison in functional gains between IRFs and SNFs can be easily made should this measure be utilized in both venues of care.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Functional_Change_Appendix-635749806898052255.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** 2016 Ad.3 **Month and Year of most recent revision:** 03, 2016 Ad.4 **What is your frequency for review/update of this measure?** Unknown, new measure Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

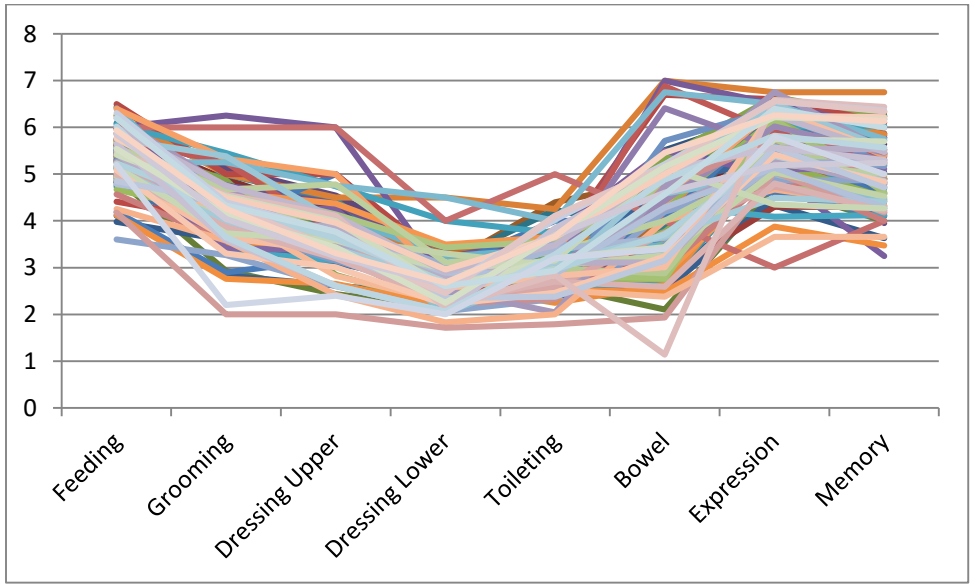
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

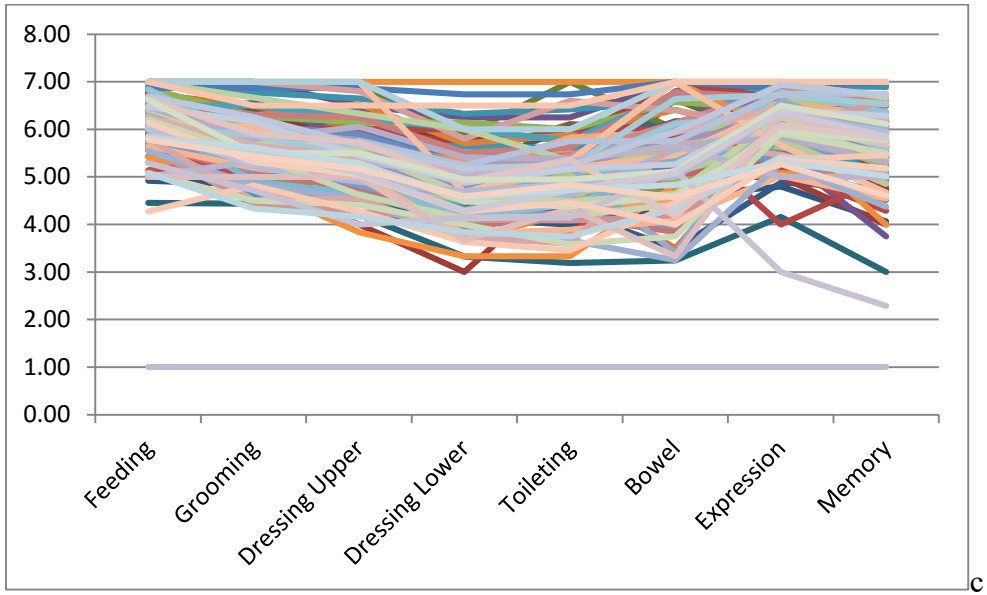
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Skilled Nursing Facilities, Change in Function: Mobility Measure for Skilled Nursing Facilities and the Change in Function: Motor Measure for Skilled Nursing Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

Self-Care Graph: Admission (Year 2015)

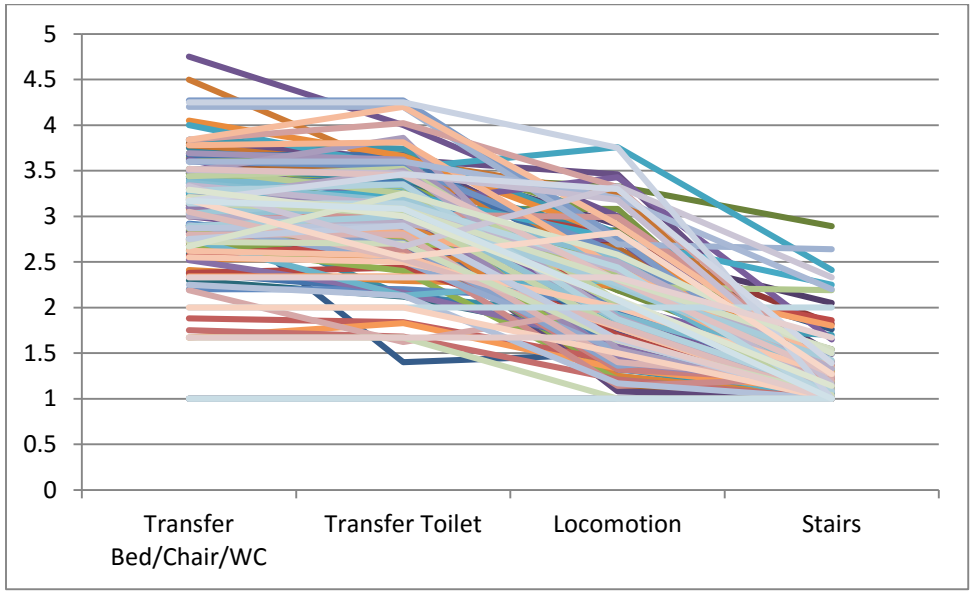


Self-Care Graph Discharge (Year 2015)

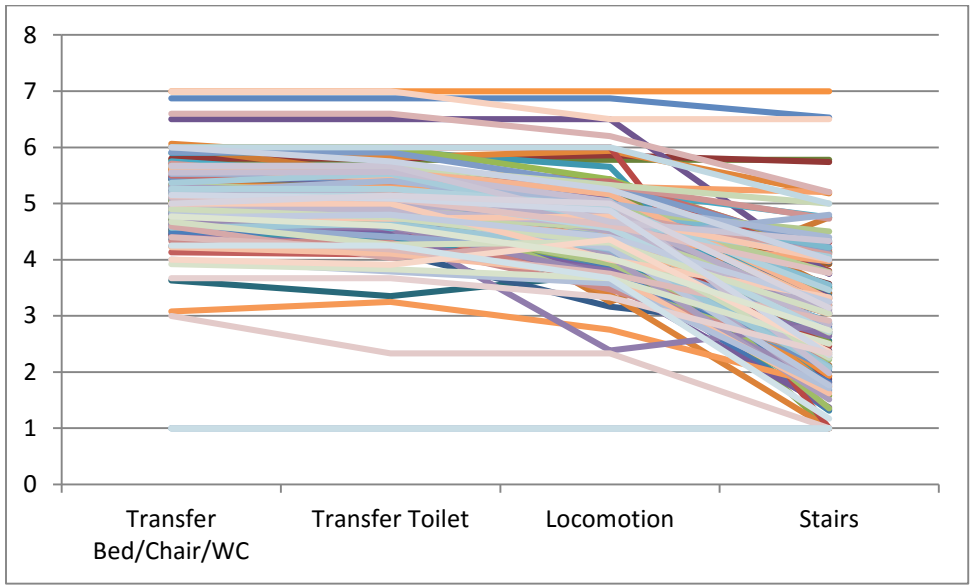


c

Mobility Graph: Admission (Year 2015)



Mobility Graph: Discharge (Year 2015)



Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the SNF venue of care.

Self-Care Mean Fit Statistics

TABLE 3.1 Self Care 8 Items ZOU018WS.TXT Mar 19 9:16 2015
INPUT: 3096 Person 8 Item REPORTED: 3094 Person 8 Item 7 CATS WINSTEPS 3.73

SUMMARY OF 2969 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.6	8.0	50.76	3.96	.96	-.1	1.02	.0
S.D.	11.5	.3	13.60	1.46	.71	1.2	.82	1.2
MAX.	55.0	8.0	87.04	10.90	6.32	5.4	8.33	6.2
MIN.	8.0	3.0	11.87	3.00	.05	-3.9	.05	-3.7

REAL RMSE	4.60	TRUE SD	12.80	SEPARATION	2.78	Person RELIABILITY .89
MODEL RMSE	4.22	TRUE SD	12.93	SEPARATION	3.06	Person RELIABILITY .90
S.E. OF Person MEAN = .25						

MAXIMUM EXTREME SCORE: 50 Person
MINIMUM EXTREME SCORE: 75 Person
LACKING RESPONSES: 2 Person

SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.2	8.0	50.33	4.59				
S.D.	12.4	.3	16.71	3.40				
MAX.	56.0	8.0	100.06	19.89				
MIN.	8.0	3.0	-.06	3.00	.05	-3.9	.05	-3.7

REAL RMSE	5.99	TRUE SD	15.60	SEPARATION	2.61	Person RELIABILITY .87
MODEL RMSE	5.71	TRUE SD	15.70	SEPARATION	2.75	Person RELIABILITY .88
S.E. OF Person MEAN = .30						

Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93

Mobility Mean Fit Statistics

```
                                                    05-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       13.8       3.7       31.44     4.51       .94    -.3    .94    -.2 |
| S.D.        6.2        .5       16.49     1.26      1.27    1.4   1.34    1.2 |
| MAX.       27.0       4.0       87.88     9.51      9.90    5.8   9.90    8.5 |
| MIN.        2.0       1.0        8.08     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89   |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92   |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       12.2       3.7       26.70     5.88                              |
| S.D.        6.9        .6       19.75     3.22                              |
| MAX.       28.0       4.0       99.95    13.79                              |
| MIN.        1.0       1.0         .02     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87   |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88   |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

```
TABLE 3.1 All Facilities 12 items                ZOU439WS.TXT  Mar 19  9:43 2015
INPUT: 3096 Person  12 Item  REPORTED: 3094 Person  12 Item  7 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 3013 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      49.2      11.6       45.63    2.83       .99    -.1   1.06     .0 |
| S.D.      17.6        .7       12.31     .98       .67    1.4    .91    1.4 |
| MAX.      83.0      12.0       88.22    9.85      5.13    5.2   9.90    7.7 |
| MIN.      10.0       4.0       10.53    2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   3.30 TRUE SD   11.86  SEPARATION  3.59  Person RELIABILITY  .93 |
|MODEL RMSE   2.99 TRUE SD   11.94  SEPARATION  3.99  Person RELIABILITY  .94 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      7 Person
  MINIMUM EXTREME SCORE:     74 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      48.4      11.7       44.66    3.21                                |
| S.D.      18.3        .7       14.26    2.51                                |
| MAX.      84.0      12.0      100.06   17.81                                |
| MIN.      10.0       4.0        -.05    2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.30 TRUE SD   13.59  SEPARATION  3.16  Person RELIABILITY  .91 |
|MODEL RMSE   4.07 TRUE SD   13.66  SEPARATION  3.36  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .26                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95
```

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2774
Measure Title: Functional Change: Change in Mobility Score for Skilled Nursing Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a
Brief Description of Measure: Change in rasch derived values of mobility function from admission to discharge among adult short term rehabilitation skilled nursing facility patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 4 mobility items:Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
Developer Rationale: The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The mobility measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part collecting data on these items. Currently, more than 150 SNFs are collecting data on these items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

Numerator Statement: Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.
Denominator Statement: Facility adjusted adjusted expected change in rasch derived values, adjusted at the Skilled Nursing Facility Case Mix Group level.
Denominator Exclusions: Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1aEvidence.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. They provide a list of 9 peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in SNFs.
- In addition, they provide summaries/abstracts from three articles that support the following: *The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay.*
- The items in the mobility measure are: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- According to the developer, "The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers."
- This is a new measure, but UDSMR has been collecting data on the FIM instrument for 20 years, so they are able to report on trends. Almost half (48%) of facilities are below expectation in 2014:

Year	2010	2011	2012	2013	2014
Mobility Change Average (Rasch)	20.6	21	21	20.9	21.2
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation	71	72	72	69	79
Number of Facilities below Expectation	57	69	75	74	75
Percent of Facilities at or above Expectation	55.5%	51.1%	49.0%	48.3%	51.3%

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the mobility measure for the years 2010 to 2014. Information is not provided on whether the differences are statistically significant, however, the data provides information on factors for consideration in assessing variation and impact on various populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**pass

**Data was supplied for evidence of the FIM tool for successful rehab. The components of the tool haven't been studied separately for outcomes specific to the mobility components being specified in this measure. It would also help to know the intended use of these components in the absence of the assessment of the self-care components. Can they be used independently to predict the level of independence or services/devices needed? I was also wondering about the timeframe for the measure being 12 months. Most short term stays are much shorter, weeks to a few months and I would be interested in knowing how the 12 month timeframe was decided.

**The measure will evaluate three health outcomes (Transfers, locomotion, and stairs) - the results will provide data which can be used by the providers to access progress in care. The provider is able to make adjustments to improve care based on the measure results.

Note - this is a provider reported measure, not a PRO

1b. Performance Gap

Comments:

**I would rate this as high, rather than moderate. If nearly half of the facilities are performing below expected level, there is ample room for improvement.

**There was data provided that looked at variation and opportunity. About 50% of the facilities are at their expected performance.

SDS data was provided but no interpretation as to whether there is any statistical differences.

**The developer states the current measures being used do not adequately address patient functional status - thus resulting in an inability to report the quality of the facilities to CMS/payers. Not sure if this is an adequate gap that warrants a new measure???

1c. PRO-PM

Comments:

**n/a Not a PRO

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, MDS data, and SNF CMG codes (case mix group)

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:
  - Average change in rasch derived mobility functional score from admission to discharge at the facility level for short term rehabilitation patients, over Facility adjusted expected change in rasch derived

values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.
    - Average is calculated as (sum of change at the patient level/total number of patients).
- The calculation algorithm is included.
- Patients under age 18 and patients who died in the SNF are excluded.
- A data dictionary is included.
- The measure is stratified by risk category.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented
- This measure uses a subset of the FIM, so a Rasch analysis was conducted to test:
    - the psychometric properties of the subset of 12 items within the three venues of post-acute care, IRFs, LTACs, and SNFs.
    - The measure reliability at both the person and item level
    - to determine the fit of each item within the measure (4 items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) through infit and outfit statistics and item specific correlations.
- Internal consistency of the critical data elements was demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but is working to do so prior to the in-person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities (perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities.* UDSMR has indicated they are working to fulfill these data needs.

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.82 to .90. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm : TBD upon submission of the facility performance score level analysis
Precise specifications – yes (box 1) -> empirical testing of data elements (box 2) -> TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity Testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Developers used concurrent validity of the FIM total score (all 18 items) with the FIM mobility score: the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (mobility items).
- Predictive validity of the mobility score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:

- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for SNFs are .659 ($p < 0.001$) at admission and .787 ($p < 0.001$) at discharge. For predicative validity of functional gain, SNFs scored 0.615 ($p < 0.001$), which is considered acceptable and for discharge disposition the C-statistic is 0.79.
- For SNFs, the r-squared values at admission were 0.454 and at discharge 0.707 for functional gain.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the facility were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developer states the following risk adjustment method: To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same SNF-CMG are expected to have similar resource utilization needs and similar outcomes.

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Skilled Nursing Facility Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *No information is provided on risk adjustment for SDS factors. Do you think the measure should include SDS factors in the risk adjustment? Why or why not?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provided additional information in an addendum, including "graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers".

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- 2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment, r-squared statistic seems moderate)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**One of the studies referenced did look at ADL's and mobility and concluded the patterns of functional change differed for these domains and for specific groups of patient - this was at the patient level, not facility level (Latham et al).
Are there different components of the FIM tool that are more predictive of improvement/discharge potential?
**(n/a due to new info being submitted by developed)

2a2. Reliability Testing

Comments:

**Internal consistency for the facilities was demonstrated.
The assessment of reliability across facilities showed the ICC of a negative value with a high p value - it might be helpful to have further explanation of those results.
**(n/a due to new info being submitted by developed)

2b.2 Validity Testing

Comments:

**Facility testing was done. A number of facilities were included.
An observation is that the number of facilities tested increased from 2010 to 2014 from 128 to 154 while the case count decreased from 26472 to 21629. It would help to know the reason for this as it would seem if the number of facilities increased the case count would as well.
A correlation of the mobility components to the overall FIM was done.
**Agree with preliminary Moderate rating for validity

2b3.-2b7. Exclusions Analysis

Comments:

**I would welcome some explanation from the developer about the SNF Case Mix Grouping used to risk adjustment scores. Is the SNF-CMG a national adjuster used by all SNFs or one that USDMR has developed for its own users?
**Exclusions include patients less than 18 and patients that died. A comment is made that patients who don't have complete data are deleted from the measure. The impact to the measure isn't clear - does this happen in only a few

facilities, in only a few patients etc. If there is variation in the % of patients being deleted or the types of patients being deleted this could affect results.
**Agree with preliminary Moderate rating for validity.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- Commercial use requires a license agreement and has a fee. The developer reports the following:
  - The Functional Change: Change in Motor Score form (this form includes the items for the mobility measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**Developer states tool used to generate the data for this measure is copyrighted but available free for "internal reporting or audit." From a consumer perspective, it is desirable that the facility scores be publicly available for choice purposes. If a facility makes its score publicly available or submits it to another entity (not USDMR) for use in public reporting, would the facility have to pay a licensing fee?
**The tool can be used within the day to day practices. Training needs to be done to use the tool along with ongoing assessment of the skills of the user.
The tool and measurement are part of a subscription to UDSMR.
It isn't clear if this measure could be done outside of that structure if a facility wanted to purchase only the FIM tool.
**Data elements are collected during care and available electronically. The only concern is related to the licensing; does this explain the preliminary rating for feasibility of Moderate?

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

- The measure is currently used for internal reporting and national benchmarking by SNFs who subscribe to the UDSMR software/outcomes reporting.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results

- New measure – not available. While a new measure to NQF, the developer does provide trending data for the rasch derived scores back to 2010:

Year	2010	2011	2012	2013	2014
Mobility Change Average (Rasch)	20.6	21	21	20.9	21.2
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation	71	72	72	69	79
Number of Facilities below Expectation	57	69	75	74	75
Percent of Facilities at or above Expectation	55.5%	51.1%	49.0%	48.3%	51.3%

Unexpected findings (positive or negative) during implementation

- None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**Moderate - easy for the facilities already using the measure and the FIM tool, hard for those facilities not already using the FIM tool as it requires staff training in order to assure consistency in patient evaluations.
**The measure is being proposed for use in accountability programs. An unintended consequence could be longer length of stays to allow patients to have a higher score. This could lead to increased costs. Is there the ability to have a balancing measure that would look at Ave LOS or costs of treatment for the level of improvement obtained?
**If this is a conflicting measure, does it make it less likely that it can/will be used? (Competes with 2612: CARE) Currently it is used by UDSMR subscribers for benchmarking; there are plans to use for public reporting in the future.

Criterion 5: Related and Competing Measures

Related or competing measures

- This measure is competing with 2612 : CARE: Improvement in Mobility

Harmonization

- None

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Mobility Score for Skilled Nursing Facilities

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1. This is a measure of:

Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

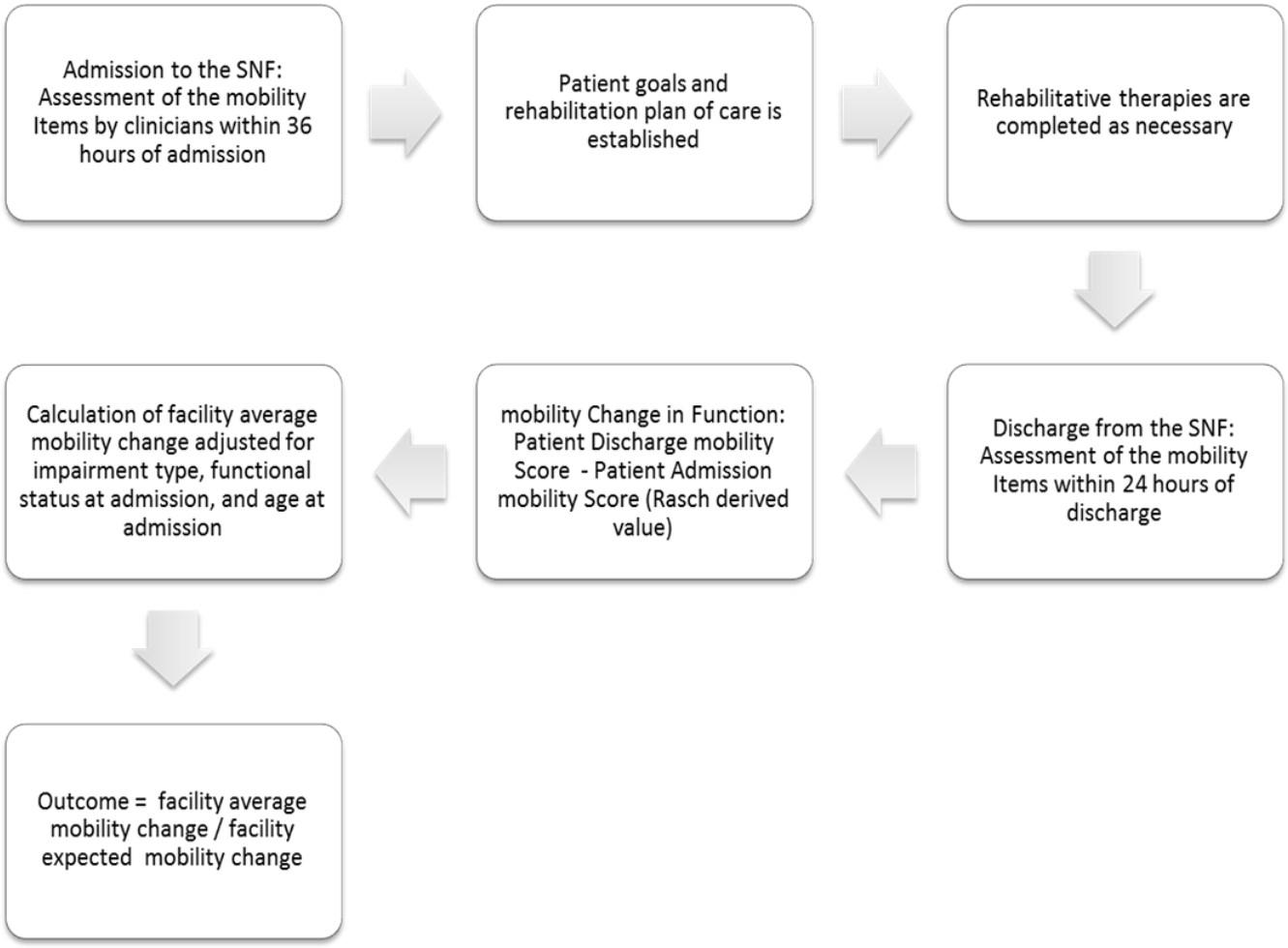
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Skilled Nursing Facilities (SNFs) are one part of a multi-level post-acute care continuum. Two different types of patients are admitted to SNFs; those meant to live in the facility, and those meant to receive short-term rehabilitation. The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are over 150 SNFs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for short term rehabilitation patients seen at SNFs. To date, the mobility measure has not been reported on as a stand-alone measure. However, the items of the mobility measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument. The mobility measure is intended to be administered within 24 hours of the patient's admission to the IRF and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the mobility measure are as follows: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for SNF short term rehabilitation patients:



While the mobility measure is new, UDSMR has been a data repository for the FIM instrument used in SNFs, of which the items of the mobility measure are nested within for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the mobility measure for the years 2010 to 2014.

Year	2010	2011	2012	2013	2014
Mobility Change Average (Rasch)	20.6	21	21	20.9	21.2
Case Count	26472	26654	26927	25620	21629
Number of Facilites at or above Expectation	71	72	72	69	79
Number of Facilities below Expectation	57	69	75	74	75
Percent of Facilities at or above Expectation	55.5%	51.1%	49.0%	48.3%	51.3%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the mobility measure for the years 2010 to 2014:

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2010		2011		2012		2013		2014	
	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)
Gender										
Male	7,668	20.9	7,705	21.3	7,617	21.4	6,489	21.3	5,100	21.7
Female	13,768	21.3	13,730	21.4	13,061	21.6	10,362	21.7	8,204	21.6
Ethnicity										
White	14,461	21.2	14,422	21.0	13,586	21.5	9,766	21.8	8,014	21.0
Black	2,073	21.6	2,273	23.0	1,997	22.4	1,609	23.1	1,453	23.4
Hispanic	370	22.9	400	22.9	353	23.8	216	22.5	140	23.4
Other Ethnicity	9,568	19.5	9,559	20.4	10,991	20.1	14,029	20.1	12,022	21.1
Payer Source										
Medicare	18,658	20.7	19,261	20.9	19,898	21.1	18,842	21.2	15,577	21.5
Medicaid	669	16.2	525	21.6	566	23.5	519	23.8	514	24.3
Commercial	1,826	21.9	2,032	21.4	2,052	21.1	2,247	19.5	1,799	20.7
Blue Cross	1,168	25.2	845	25.4	876	25.5	999	23.5	526	22.4
Other Payer	4,151	19.2	3,991	20.0	3,535	19.1	3,013	18.7	3,213	19.6
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	3,481	20.3	3,310	19.0	3,784	18.5	3,539	17.8	3,437	18.1
P02 (NY, NJ, PR)	9,099	23.0	7,581	22.3	6,031	23.7	6,290	23.6	4,426	23.8
P03 (PA, WV, VA, DE, MD, DC)	1,793	20.1	1,489	22.0	1,565	24.7	1,721	23.2	1,198	23.1
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	8,057	18.0	7,542	21.9	7,401	20.1	8,759	19.7	7,405	21.0
P05 (MN, WI, IL, IN, MI, OH)	3,728	21.4	3,290	20.9	3,313	21.6	4,289	21.1	4,907	21.3
P06 (NM, OK, AR, LA, TX)	29	20.8	2,015	17.4	2,685	20.0	383	22.0	0	-
P07 (NE, IA, KS, MO)	285	17.7	1,381	18.1	2,124	18.7	639	21.0	135	17.9
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	33	19.4
P09 (CA, NV, AZ, HI)	0	-	46	22.2	24	23.9	0	-	88	17.5
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the mobility measure is a new measure and has not been used as a stand-alone tool. However all of the items within the measure are included in a larger instrument (the FIM instrument) which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the mobility measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and use of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Skilled Nursing Facilities:

1. Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.
2. Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.

3. Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.

4. Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

5. Munin MC, Begley A, Skidmore ER, Lenze EJ. Influence of rehabilitation site on hip fracture recovery in community-dwelling subjects at 6-month follow-up. *Archives of physical medicine and rehabilitation.* Jul 2006;87(7):1004-1006.

6. Munin MC, Seligman K, Dew MA, et al. Effect of rehabilitation site on functional recovery after hip fracture. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):367-372.

7. Nelson DL, Melville LL, Wilkerson JD, Magness RA, Grech JL, Rosenberg JA. Interrater reliability, concurrent validity, responsiveness, and predictive validity of the Melville-Nelson Self-Care Assessment. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association.* Jan-Feb 2002;56(1):51-59.

8. Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Archives of physical medicine and rehabilitation.* Oct 1996;77(10):1056-1061.

9. Vincent KR, Vincent HK. A multicenter examination of the Center for Medicare Services eligibility criteria in total-joint arthroplasty. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Jul 2008;87(7):573-584.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.

OBJECTIVE: To examine the outcomes of patients with varying levels of cognitive impairment who received rehabilitation in skilled nursing facilities (SNFs). DESIGN: A retrospective analysis of the records of people admitted to SNFs for rehabilitation. SETTING: Seven SNFs in Colorado. PARTICIPANTS: Community-dwelling persons (N=7159), 65 years of age and older, admitted for rehabilitation after a hospitalization or decline in function between May 1998 and May 2002. Interventions Not applicable. MAIN OUTCOME MEASURES: Cognitive impairment was assessed using a 4-level categorization of the FIM instrument cognitive score at admission. Functional gain was measured using the FIM. Community discharge was measured as the proportion of patients discharged to home, board and care, or assisted living facility. Rehabilitation progress was measured as the number of FIM points gained per day. RESULTS: Significant functional gains were made during rehabilitation in motor and cognitive FIM scores, regardless of cognitive impairment. The most cognitively impaired patients required more rehabilitation intervention, achieved less FIM gain, and were less likely to be discharged to the community. The strongest predictors of FIM gain were the amount of therapy hours and admission cognitive FIM score. The strongest predictors of discharge to the community were the discharge total FIM score and age. The strongest predictors of adequate rehabilitation progress were medical complexity and admission cognitive FIM score. CONCLUSIONS: Patients with cognitive impairment were able to recover function with rehabilitation intervention. Patients with a more serious cognitive impairment received more rehabilitation intervention than patients with less impairment. Outcomes were predicted by

admission and rehabilitation measures that were qualitatively different from other discharge outcomes. Health care professionals need to consider these factors as they create a rehabilitation plan of care for patients with cognitive impairment.

Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.

OBJECTIVES: To document patient, program characteristics, and therapy service provision in subacute rehabilitation across 3 types of facilities that provide subacute rehabilitation, to examine the determinants of therapy intensity, and to evaluate the contribution of rehabilitation services to functional gains. DESIGN: A retrospective study linking administrative billing data and patients' functional assessment records. SETTING: Twenty facilities part of the Uniform Data System for Medical Rehabilitation (UDSMR) subacute database PARTICIPANTS: A total of 1976 billing records of patients with stroke, orthopedic, and debility impairments, discharged in 1996 and 1997, were retrieved and linked with the FIM trade mark instrument ratings from UDSMR subacute database. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: Total therapy intensity and Rasch-transformed FIM domain gains (ie, gains in self-care, mobility, cognition). RESULTS: Therapy intensity was mostly determined by impairment and facility type, although variances explained by the predictors were small. Patients in all 3 impairment groups made functional gains; gains were related weakly, although significantly, to therapy intensity and rehabilitation duration after controlling for other variables. CONCLUSIONS: The provision of rehabilitation therapies varied across facilities. Skilled nursing facilities with subacute rehabilitation units tended to provide more therapies than subacute units in acute or rehabilitation hospitals.

Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.

OBJECTIVE: To examine the relation between therapy intensity, including physical therapy (PT), occupational therapy (OT), and speech and language therapy (SLT), provided in a skilled nursing facility (SNF) setting and patients' outcomes as measured by length of stay (LOS) and stage of functional independence as measured by the FIM instrument. DESIGN: A retrospective analysis of secondary data from an administrative dataset compiled and owned by SeniorMetrix Inc. SETTING: Seventy SNFs under contract with SeniorMetrix health plan clients. PARTICIPANTS: Patients with stroke, orthopedic conditions, and cardiovascular and pulmonary conditions (N=4988) covered by Medicare+Choice plans, and admitted to an SNF in 2002. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: LOS and improvement in stage of independence in the mobility, activities of daily living (ADLs), and executive control domains of function as determined by the FIM instrument. RESULTS: Higher therapy intensity was associated with shorter LOS (P <.05). Higher PT and OT intensities were associated with greater odds of improving by at least 1 stage in mobility and ADL functional independence across each condition (P <.05). The OT intensity was associated with an improved executive control stage for patients with stroke, and PT and OT intensities were associated with improved executive control stage for patients with cardiovascular and pulmonary conditions (P <.05). The SLT intensity was associated with improved motor and executive control functional stages for patients with stroke (P <.05). Therapy intensities accounted for small proportions of model variances in all outcomes. CONCLUSIONS: Higher therapy intensity was associated with better outcomes as they relate to LOS and functional improvement for patients who have stroke, orthopedic conditions, and cardiovascular and pulmonary conditions and are receiving rehabilitation in the SNF setting.

Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

OBJECTIVE: To examine the rate of functional change in 2 domains, activities of daily living (ADLs) and mobility, over 2 time periods during hip fracture rehabilitation. DESIGN: Retrospective analysis of data contained in an administrative dataset. SETTING: Seventy skilled nursing facilities (SNFs). PARTICIPANTS: People (N=351) receiving rehabilitation in SNFs from March 1998 to February 2003 after hip fractures. INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURE: Rate of change in scores in the ADL and mobility domains of the FIM instrument during 2 time intervals of rehabilitation. RESULTS: The rate of functional change across 2 time intervals was constant for mobility (mean change in FIM points per day, .46 vs .49), but declined in the second time period for ADLs (mean change in FIM points per day, .55 vs .41). Executive function, length of stay (LOS), and medical complexity were related to rate of change in mobility, and baseline ADLs, executive function, living setting, and LOS were related to rate of change in ADLs. There was an interaction between rehabilitation phase and baseline mobility. People with lower baseline mobility had an increased rate of change during the second interval (mean change in FIM points per day, .41 vs .55), whereas those with higher baseline mobility had a decreased rate of change (mean change in FIM points per day, .50 vs .43). CONCLUSIONS: The pattern of functional change over time

differed for ADL and mobility domains, and for specific groups of patients. The results have implications for goal setting and discharge planning.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2774
De.2. **Measure Title:** : Functional Change: Change in Mobility Score for Skilled Nursing Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a
De.3. **Brief Description of Measure:** Change in rasch derived values of mobility function from admission to discharge among adult short term rehabilitation skilled nursing facility patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 4 mobility items:Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
1b.1. **Developer Rationale:** The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The mobility measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part collecting data on these items. Currently, more than 150 SNFs are collecting data on these items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

S.4. **Numerator Statement:** Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.
S.7. **Denominator Statement:** Facility adjusted adjusted expected change in rasch derived values, adjusted at the Skilled Nursing Facility Case Mix Group level.
S.10. **Denominator Exclusions:** Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Mobility_SNF-635950324345357206.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The mobility measure is constructed by utilizing items which are presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part collecting data on these items. Currently, more than 150 SNFs are collecting data on these items for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.

We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Please see Measure Evaluation Form for data over time

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see Measure Evaluation Form for disparities data

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer, Cardiovascular, Musculoskeletal, Neurology, Pulmonary/Critical Care

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment**:** NQF_Submission_Mobility-635749898391586121.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in
the SNF. The numerator is the average change in rasch derived mobility functional score from admission to discharge for each
patient at the facility level, including items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level for all items (Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted adjusted expected change in rasch derived values, adjusted at the Skilled Nursing Facility Case Mix Group level.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in
the SNF. Impairment type is defined as the primary medical reason for the SNF short term rehabilitation stay (such as stroke, joint
replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum 4 items (Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) at the facility level. Age is the age
of the patient at the time of admission to the SNF. The denominator is meant to reflect the expected Mobility functional change score at the facility, if the facility had the same distribution of SNF-CMGs (based on impairment type, functional status at admission, and age at admission). This adjustment procedure is an indirect standarization procedure (observed facility average/expected
facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Excluded in the measure are patients who died in the SNF or patients less than 18 years old.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through the MDS.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the SNF-CMGs in the excel file provided.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standardization procedure (observed facility average/expected facility average). The
numerator is the facility´s average mobility functional change score. The denominator is meant to reflect the expected Mobility functional change score at the facility, if the facility had the same distribution of SNF-CMGs(impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all short term rehabilitation patients during the assessment time frame (12 months).
2. Exclude any patients who died in the SNF.
3. Exclude any patients who are less than 18 at the time of admission to the SNF.
3. Calculate the total mobility change score for each of the remaining patients (sum of change at the patient level for all items
(Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in the excel file).
5. Calculate the average rasch derived mobility change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average mobility
change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average mobility change score/facility´s national expected mobility
change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.

Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry

S.24. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the http://share.qualityforum.org/Projects/person_and_family_care/CommitteeDocuments/Functional%20Change_Change%20in%20Mobility%20Score%20for%20Skilled%20Nursing%20Facilities/PFCC3_2774_Functional_Change_Appendix.pdf.

S.25. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Mobility_SNF-635950324531469978.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Mobility Score for Skilled Nursing Facilities
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically

significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® (“FIM”)instrument data from inpatient rehabilitation facilities, long term acute care facilities, and skilled nursing facilities from the Uniform Data System for Medical Rehabilitation. The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the mobility measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2002-2013 were used in examining the data trends over time using the mobility measure and patient outcomes of inpatient rehabilitation

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: patient level/aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term acute care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument(the tool used for this measure) is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 4 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for the skilled nursing facility venue of care. However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Subsequently, this measure is being submitted for all three venues of care. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our mobility measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

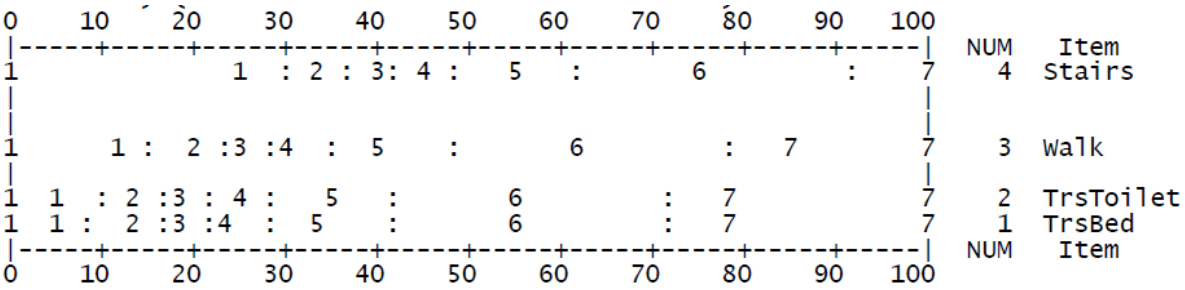
Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (4 items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistant compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.82 to 0.90. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure, we set the Rasch scale at 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.



The ruler shows that the easiest item is Transfers: Bed/Chair/Wheelchair, and the hardest Stairs and that the distances between a level 1 and 2 and 5, 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 4 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
4	.00E	12.48	13	26.23	3.76	22	50.27	5.85
5	8.05	6.65	14	28.41	3.94	23	55.99	6.63
6	12.45	4.65	15	30.76	4.06	24	62.97	7.09
7	15.07	3.91	16	33.17	4.04	25	70.32	7.08
8	17.10	3.59	17	35.50	3.95	26	77.95	7.52
9	18.91	3.46	18	37.76	3.93	27	87.92	9.24
10	20.65	3.45	19	40.08	4.07	28	100.00E	13.82
11	22.41	3.50	20	42.69	4.42			
12	24.25	3.61	21	45.94	5.04			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As indicated previously, the reliability of the FIM instrument is well known. The results of the analysis for the measure proposed show the reliability holds even when looking at a subset of FIM instrument items.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the mobility measure with the FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new mobility measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (mobility items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the mobility measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM® instrument (the gold standard)), and likelihood of discharge to the community setting Linear regression was used to determine functional change, whereas the change in mobility was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission mobility total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the mobility measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was correlated with the FIM total, 0.671 ($p < 0.001$) and 0.768 ($p < 0.001$), respectively. The correlations remained significant within each venue of care; IRFs, 0.605 ($p < 0.001$) and 0.847 ($p < 0.001$); LTACs, 0.711 ($p < 0.001$) and 0.764 ($p < 0.001$); SNFs, 0.659 ($p < 0.001$) and 0.787 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM totals, the r-square values ranged from respectable for admission FIM total, to high for discharge FIM total, 0.512 and 0.706, respectively. The values remained similar at the venue specific level as well; IRFs, 0.400 and 0.676; LTACs, 0.540 and 0.707; SNFs, 0.454 and 0.707.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was acceptable, 0.615 ($p < 0.001$). In addition, by venue, the correlations remained acceptable; IRFs, 0.598 ($p < 0.001$); LTACs, 0.665 ($p < 0.001$); SNFs, 0.611 ($p < 0.001$). The linear regression showed acceptable r-squared values as well; all venues, 0.506; IRFs, 0.438; LTACs, 0.559; SNFs, 0.486.

Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in our measure has good predictive ability for discharge setting (community), with a C-statistic of 0.79. By venue, the results are similar; IRFs, 0.78; LTACs, 0.77; SNFs, 0.77.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results show good validity across all analyses. The r-square values were all consistent around 0.5 – 0.6, meaning that the percent of variance explained in the dependent variables by our measure were all more than 50%. Considering we are testing the correlation between 4 items of an 18 item scale, these r-squared values are quite good. In addition, the predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients that died in the post-acute care setting (an unanticipated outcome) and patient aged 18 years and older, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 1 **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. SNF-CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM instrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all SNF-CMGs.)

See file uploaded in S.15 for calculations.

The SNF-CMGs are groupings specific to skilled nursing facilities, although they are similar and easily comparable to the CMGs used in inpatient rehabilitation facilities.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, SNF- CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.
2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.
4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
While this is a new measure, the data collection procedure for items is in place for SNFs utilizing UDSMR software.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The Functional Change: Change in Motor Score form (this form includes the items for the mobility measure) submitted is

copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or
internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices.
Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial
gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or
distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license
agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per
commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Uniform Data System for Medical Rehabilitations http://www.udsmr.org/ Quality Improvement (Internal to the specific organization) Uniform Data System for Medical Rehabilitations www.udsmr.org

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for SNFs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for skilled nursing facilities, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

The FIM System® provides an established means of collecting rehabilitation data in a consistent manner. It allows clinicians to follow changes in the functional status of their patients from the start of rehabilitative care through discharge and follow-up.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

We are applying for initial endorsement.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
As we used existing data that has already been colected, there were no unintended negative consequences to individuals or populations identified during our testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
2612 : CARE: Improvement in Mobility

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
While the CARE items and the change in mobility items measure the same construct of functional (in)dependence, there are some key differences included in the measures, and in the measurement of the items. The mobility measure, submitted by UDS includes the following items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. The CARE items included in the measure submitted by AHCA include: : Roll left and right, Sit to lying, Lying to sitting on side of bed, Sit to stand, Chair/bed-to-chair transfer, Toilet transfer, Car transfer, Walk 10 feet, Walk 50 feet with 2 turns, Walk 150 feet, Walking 10 feet on uneven surfaces, 1 step, 4 steps, 12 steps, Pick up object. Once again there is great overlap in the items, There is great overlap between the items in the two measures, particularly in the transfer items, locomotion, and stairs. However while our measure contains only four items, the CMS measure contains 14 items. While our measure has the one locomotion item, for instance, the ACHA measure has four. Similarly, our measure contains one item for stairs, while the CMS measure contains three. This becomes burdensome on the provider to have to collect an additional 10 items and it hasn't been proven that there is additional value or specificity in the measure. Rasch analysis shows us that more items do not always mean better measurement. Finally, the UDSMS change in mobility measure is the exact same measure (same items, same rating scale, same adjustment) used in SNF, IRF and LTAC, offering consistency in measuring patient function across PAC venues, which has been an interest for PAC and is a current objective of the IMPACT ACT.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
The functional items have been collected in SNFs for over 20 years. This allows for a historical perspective of function in the SNFs that the CARE items do not allow. In addition, the these items have been used in inpatient rehabilitation facilities for over 30 years, and therefore, a comparison in functional gains between IRFs and SNFs can be easily made should this

measure be utilized in both venues of care.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Functional_Change_Appendix-635749898140419681.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 03, 2016
Ad.4 **What is your frequency for review/update of this measure?** Unknown, new measure
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

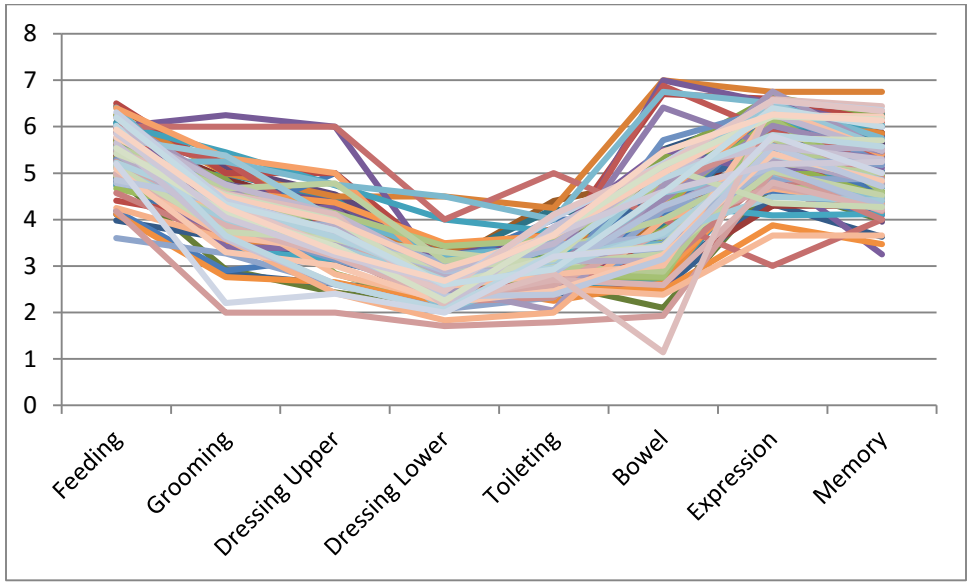
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

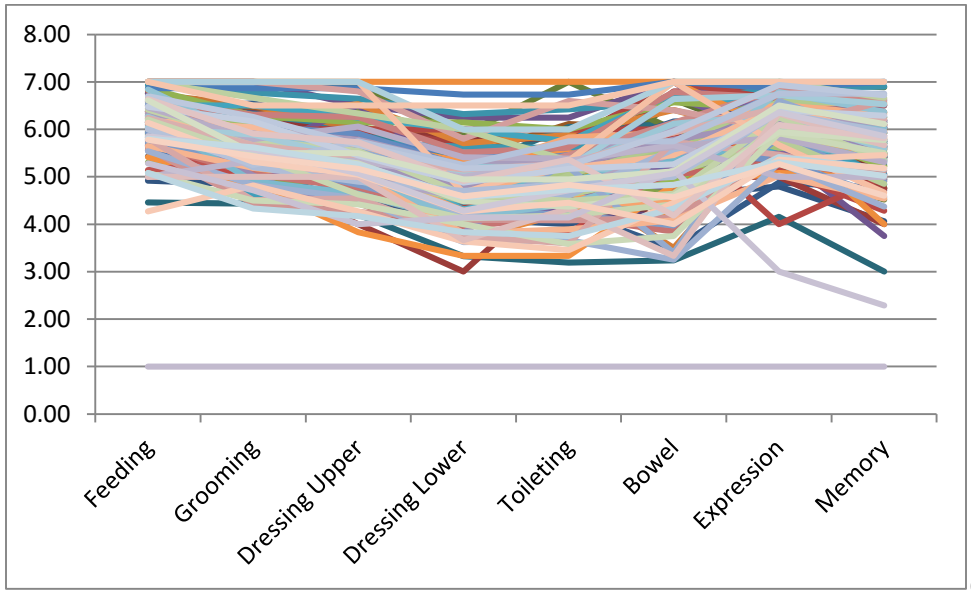
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Skilled Nursing Facilities, Change in Function: Mobility Measure for Skilled Nursing Facilities and the Change in Function: Motor Measure for Skilled Nursing Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

Self-Care Graph: Admission (Year 2015)

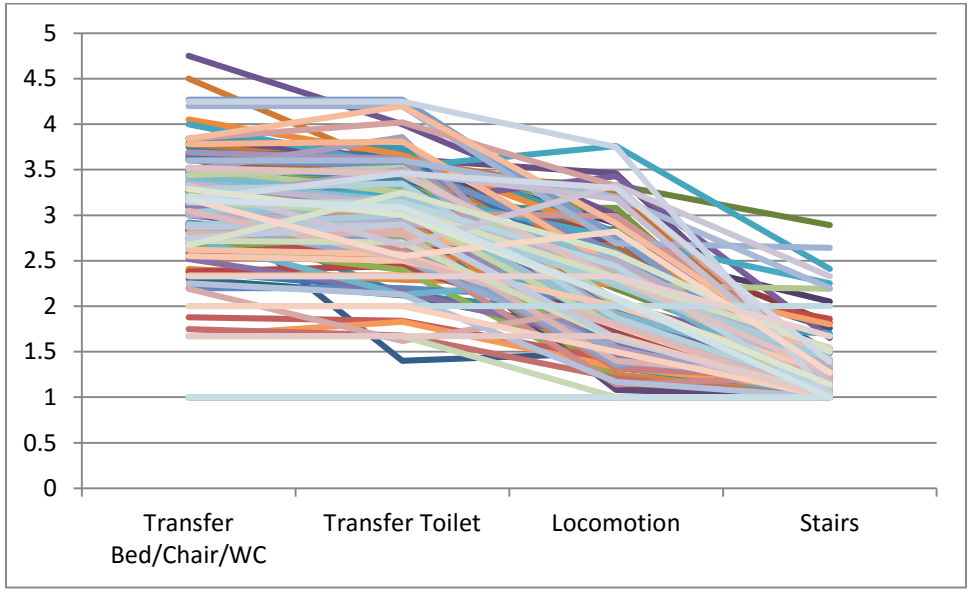


Self-Care Graph Discharge (Year 2015)

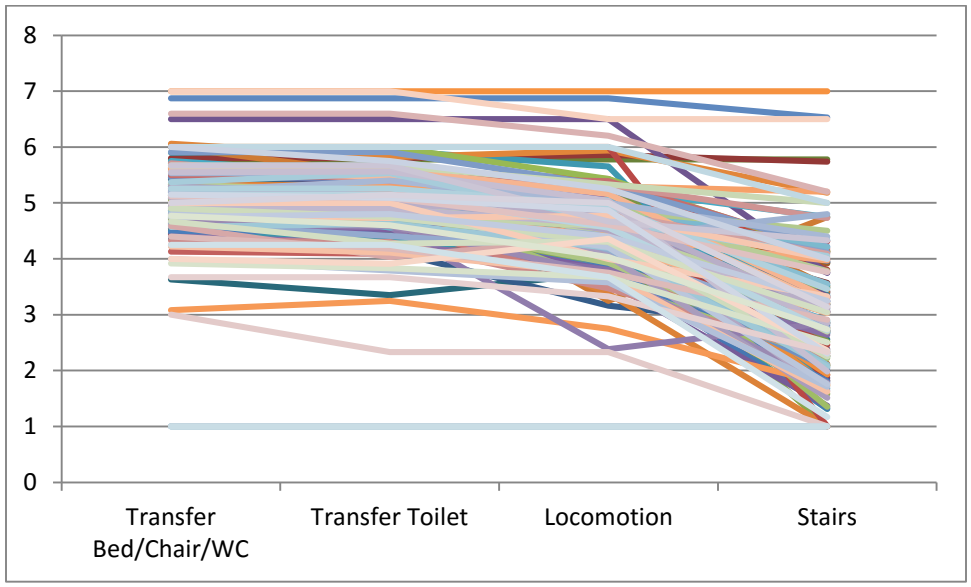


c

Mobility Graph: Admission (Year 2015)



Mobility Graph: Discharge (Year 2015)



Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the SNF venue of care.

Self-Care Mean Fit Statistics

```
TABLE 3.1 Self Care 8 Items                    ZOU018WS.TXT  Mar 19  9:16 2015
INPUT: 3096 Person  8 Item  REPORTED: 3094 Person  8 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2969 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      36.6       8.0       50.76     3.96       .96    -.1   1.02     .0 |
| S.D.      11.5        .3       13.60     1.46       .71    1.2    .82    1.2 |
| MAX.      55.0       8.0       87.04    10.90      6.32    5.4   8.33    6.2 |
| MIN.       8.0       3.0       11.87     3.00       .05   -3.9    .05   -3.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.60 TRUE SD   12.80  SEPARATION  2.78  Person RELIABILITY  .89 |
|MODEL RMSE   4.22 TRUE SD   12.93  SEPARATION  3.06  Person RELIABILITY  .90 |
| S.E. OF Person MEAN = .25                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     50 Person
  MINIMUM EXTREME SCORE:     75 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      36.2       8.0       50.33     4.59                               |
| S.D.      12.4        .3       16.71     3.40                               |
| MAX.      56.0       8.0      100.06    19.89                               |
| MIN.       8.0       3.0        -.06     3.00       .05   -3.9    .05   -3.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.99 TRUE SD   15.60  SEPARATION  2.61  Person RELIABILITY  .87 |
|MODEL RMSE   5.71 TRUE SD   15.70  SEPARATION  2.75  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .30                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93
```

Mobility Mean Fit Statistics

```
                                                    03-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       13.8       3.7       31.44     4.51       .94    -.3    .94    -.2 |
| S.D.        6.2        .5       16.49     1.26      1.27    1.4   1.34    1.2 |
| MAX.       27.0       4.0       87.88     9.51      9.90    5.8   9.90    8.5 |
| MIN.        2.0       1.0        8.08     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89 |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       12.2       3.7       26.70     5.88                              |
| S.D.        6.9        .6       19.75     3.22                              |
| MAX.       28.0       4.0       99.95    13.79                              |
| MIN.        1.0       1.0         .02     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87 |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

```
TABLE 3.1 All Facilities 12 items              ZOU439WS.TXT  Mar 19  9:43 2015
INPUT: 3096 Person  12 Item  REPORTED: 3094 Person  12 Item  7 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 3013 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      49.2      11.6       45.63     2.83       .99    -.1   1.06     .0 |
| S.D.      17.6        .7       12.31      .98       .67    1.4    .91    1.4 |
| MAX.      83.0      12.0       88.22     9.85      5.13    5.2   9.90    7.7 |
| MIN.      10.0       4.0       10.53     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   3.30 TRUE SD   11.86  SEPARATION  3.59  Person RELIABILITY  .93 |
|MODEL RMSE   2.99 TRUE SD   11.94  SEPARATION  3.99  Person RELIABILITY  .94 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      7 Person
  MINIMUM EXTREME SCORE:     74 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      48.4      11.7       44.66     3.21                               |
| S.D.      18.3        .7       14.26     2.51                               |
| MAX.      84.0      12.0      100.06    17.81                               |
| MIN.      10.0       4.0        -.05     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.30 TRUE SD   13.59  SEPARATION  3.16  Person RELIABILITY  .91 |
|MODEL RMSE   4.07 TRUE SD   13.66  SEPARATION  3.36  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .26                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95
```

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2775
Measure Title: Functional Change: Change in Motor Score for Skilled Nursing Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Brief Description of Measure: Change in rasch derived values of motor function from admission to discharge among adult short term rehabilitation skilled nursing facility patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 12 items: Feeding, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
Developer Rationale: The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing functional items presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the measure. Currently more than 150 SNFs are utilizing the items in our proposed measure for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

Numerator Statement: Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the SNF or patients who died within the SNF are excluded.
Denominator Statement: Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.
Denominator Exclusions: Patients age at admission less than 18 years old
Patients who died in the SNF.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. They provide a list of 9 peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in SNFs.
- In addition, they provide summaries/abstracts from three articles that support the following: *The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay.*
- The items in the motor score measure are: Feeding, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

According to the developer, "The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers."

This is a new measure, but UDSMR has been collecting data on the FIM instrument for 20 years, so they are able to report on trends. Almost half (46.1%) of facilities are below expectation in 2014:

Year	2010	2011	2012	2013	2014
Motor Change Average (Rasch)	11.9	12.4	12.1	12.0	12.1
Case Count	26,472	26,654	26,927	25,620	21,629
Number of Facilites at or above Expectation	72	80	77	74	83
Number of Facilities below Expectation	56	61	70	69	71
Percent of Facilities at or above Expectation	56.3%	56.7%	52.4%	51.7%	53.9%

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the motor measure for the years 2010 to 2014. Information is not provided on whether the differences are statistically significant, however, the data provides information on factors for consideration in assessing variation and impact on various populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**There is evidence that supports the need for a measure that assesses the change in function status related to motor skills between admission and discharge.

I am concerned that the metrics submitted were designed and tested largely on populations other than the one for which this is intended. A portion of the submission includes documentation that shows the cross-out of IRF and IRF-PAI notations, substituting SNF instead, suggesting that a single measure, or combination of measures, are interchangeable between IRF &

SNF populations. In addition, no adjustment appears to be included that would evaluate and control for level of care differences.

The data table showing 2010 thru 2014 data for short-term SNF stays is presented at the individual stay level and offers to provide the same data at the facility level. I would like to see the offered table.

The references provided contain data which is sometimes old (1996 & 1997), targeted diagnostic groups and do not differentiate between levels of post-acute care, and/or do not control for level of care or length of inpatient stay(s) prior to admission to the SNF level of care. This is especially troubling since the conclusion drawn from one reference states its conclusion, "The provision of rehabilitation therapies varied across facilities. Skilled nursing facilities with subacute rehabilitation units tended to provide more therapies than subacute units in acute or rehabilitation hospitals". Yet no such breakout or control if part of the submission.

**There is a clear relationship between the outcome and the care planning process within the LTC facility with a focus on sub-acute.

1b. Performance Gap

Comments:

**SNF data, including mobility, are currently part of the CMS required MDS tool. I did not see any reference to that tool or comparisons with its metrics.

CMS MDS Tool
https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/NHQIMDS30.html

AMRPA article Summary
http://www.amrpa.org/newsroom/Dobson%20DaVanzo%20AMRPA%202-page%20summary%20REVISED%203.10.14%20DATED%207.10.14.pdf

AMRPA article Full Version
http://www.amrpa.org/newsroom/Dobson%20DaVanzo%20Final%20Report%20-%20Patient%20Outcomes%20of%20IRF%20v%20%20SNF%20-%207%2010%2014%20redated.pdf

**There currently exists outcome measures that address the short term residents in a long term care facility. Not all outcomes addressed here are reported.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, MDS data, and SNF CMG codes (case mix group)

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:
  - Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients, over Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.
  - Average is calculated as (sum of change at the patient level/total number of patients).
- The calculation algorithm is included.
- Patients under age 18 and patients who died in the SNF are excluded.
- A data dictionary is included.
- The measure is stratified by risk category.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented
- This measure uses a subset of the FIM, so a Rasch analysis was conducted to test:
  - the psychometric properties of the subset of 12 items within the three venues of post-acute care, IRFs, LTACs, and SNFs.
  - The measure reliability at both the person and item level
  - to determine the fit of each item within the measure (12 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory and Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations.
- Internal consistency of the critical data elements was demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but us striving to do so prior to the in—person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities (perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they*

would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities. UDSMR has indicated they are working to fulfill these data needs.

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.94. The Cronbach Alpha reliability statistic was 0.95. Item correlations within the measure ranged from 0.65 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empirical testing of data elements (box 2): Yes→ TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Developers used concurrent validity of the FIM total score (all 18 items) with the FIM motor score: the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (motor items).
- Predictive validity of the motor score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:

- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for SNFs are .944 ($p < 0.001$) at admission and .947 ($p < 0.001$) at discharge. For predicative validity, SNFs scored 0.837 ($p < 0.001$).
- The r-squared values were all above 0.8, meaning that the percent of variance explained in the dependent were all more than 80%. For SNFs, the r-squared values at admission were 0.960 and at discharge 0.980 for functional gain.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the facility were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developer states the following risk adjustment method: To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same SNF-CMG are expected to have similar resource utilization needs and similar outcomes.

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Skilled Nursing Facility Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Specific questions on the risk-adjustment approach.*
- *Is an appropriate risk-adjustment strategy included in the measure?*

- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *No information is provided on risk adjustment for SDS factors. Do you think the measure should include SDS factors in the risk adjustment? Why or why not?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provided additional information in an addendum, including "graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers".

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**This measure seeks to evaluate the change in motor function for a patient in a short-stay SNF setting, yet uses IRF & LTAC data, combined with SNF data, to establish reliability and validity, as well as it's evaluation of the quality of SNF care. This member would suggest reproducing these numbers using all available SNF-only data.

**No inconsistencies were apparent.

2a2. Reliability Testing

Comments:

**With regard to the new information provided about the ICC data, I am unclear on the process. A split half reliability study would evaluate the internal consistency of the measure, not the difference between facilities. If the facilities were split and tested against themselves first, to show internal reliability through no significant differences being identified, and then facility-level data tested against each other suggesting significant differences between them makes more sense to me.

**The FIMS tool has a history of demonstrated reliability. Facility level data were reported.

2b.2 Validity Testing

Comments:

**Why were only 6 SNF facilities included in the analyses when they state that they have data on over a hundred? The data from the 150 SNF facilities offer data from 0.1% of the SNFs in the USA. In addition, all SNFs in the pool are voluntarily using the FIM. No evidence has been submitted that compares these SNFs to SNFs not in the database.
**The facility lever were tested.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- Commercial use requires a license agreement and has a fee. The developer reports the following:
    - The Functional Change: Change in Motor Score form (this form includes the items for the motor measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**SNF does not use the FIM tool. The FIM tool certainly provides solid IRF data, however to use the FIM requires a clinician, who has been tested and certified as competent to administer and collect data on the tool. As this would be a new tool, new measures and requires training, the burden to the system in initial & ongoing expenses may outweigh the benefits.

I also feel compelled to say that I have concerns regarding the objectivity of the measure steward given that UDSMR services, which include costs and fees for their tools and certification processes, stand to benefit financially from the integration of their measures in other health care settings.

**All measures should be available electronically. MDS does no result in data collection for all these measures. It would appear that the facility would need to purchase the ability to use the FIMS tool.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

- The measure is currently used for internal reporting and national benchmarking by SNFs who subscribe to the UDSMR software/outcomes reporting.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results

- New measure – not available. While a new measure to NQF, the developer does provide trending data for the rasch derived scores back to 2010:

Year	2010	2011	2012	2013	2014
Motor Change Average (Rasch)	11.9	12.4	12.1	12.0	12.1
Case Count	26,472	26,654	26,927	25,620	21,629
Number of Facilites at or above Expectation	72	80	77	74	83
Number of Facilities below Expectation	56	61	70	69	71
Percent of Facilities at or above Expectation	56.3%	56.7%	52.4%	51.7%	53.9%

Unexpected findings (positive or negative) during implementation

- None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**I would suggest that given the challenges with feasibility and concern over the development of a new revenue stream for the measure steward, which is a direct conflict of interest, other options may provide an equally valuable cross cutting measure by adding non-proprietary items to the current MDS tool.

**Currently not publicly reported however a some short term measures are currently reported on Nursing Home Compare.

Criterion 5: Related and Competing Measures

Related or competing measures

None listed, however, this measure is the "parent" to the mobility and self-care measures that have been identified as competing with measures: 2612: CARE Improvement in Mobility and 2613: Care Improvement in Self-Care

Harmonization
None

Pre-meeting public and member comments

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Motor Score for Skilled Nursing Facilities
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

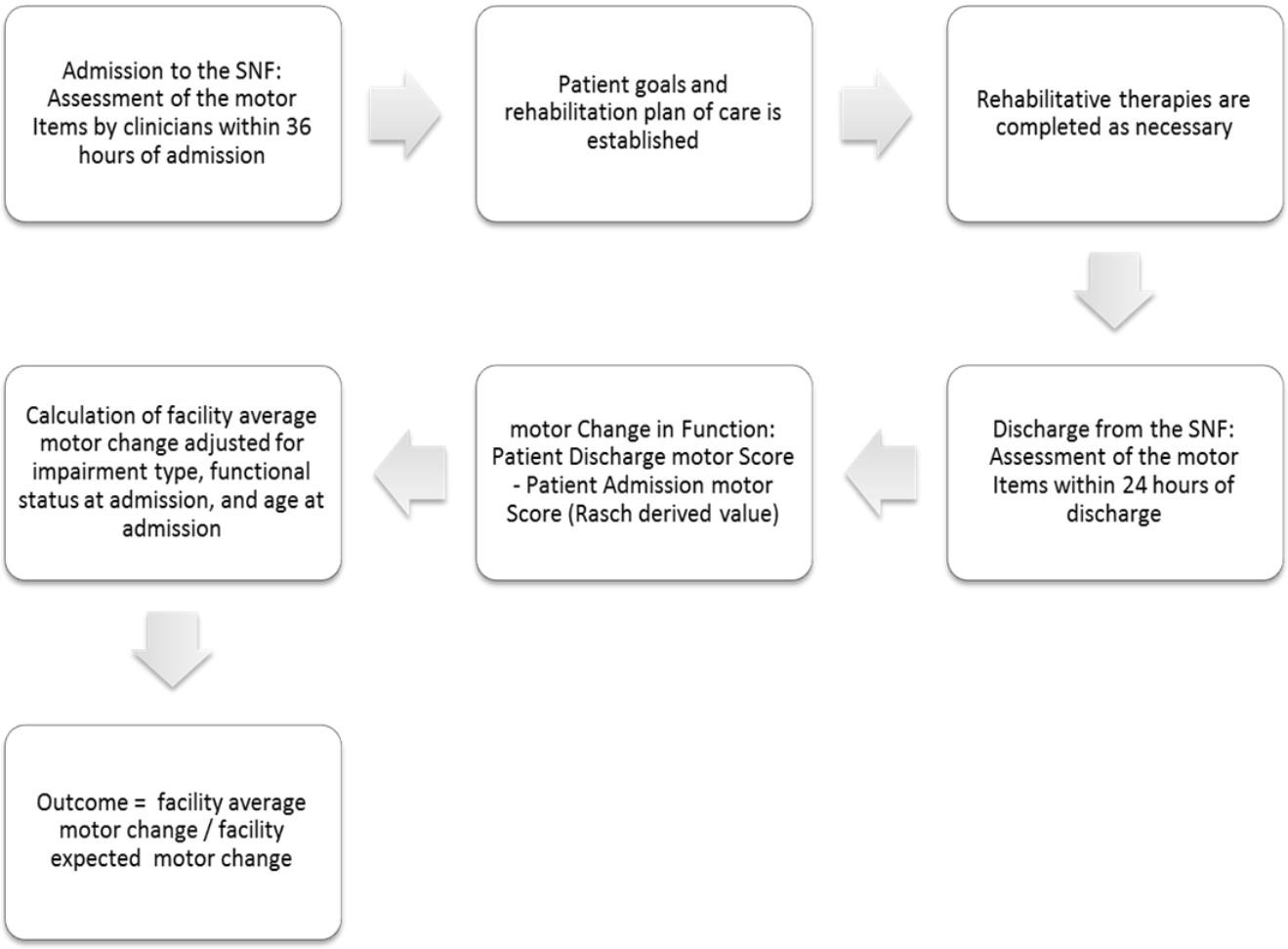
☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Skilled Nursing Facilities (SNFs) are one part of a multi-level post-acute care continuum. Two different types of patients are admitted to SNFs; those meant to live in the facility, and those to receive short-term rehabilitation. The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in

inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are over 150 SNFs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for short term rehabilitation patients seen at SNFs. To date, the motor measure has not been reported on as a stand-alone measure. However, the items of the motor measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument.. The motor measure is intended to be administered within 24 hours of the patient's admission to the SNF and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the motor measure are as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs). All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for SNF short term rehabilitation patients:



UDSMR has been a data repository for the FIM instrument among SNF patients, of which the items of the motor measure are nested within for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the self-care measure for the years 2010 to 2014 for short term skilled nursing facility patients:

Year	2010	2011	2012	2013	2014
Motor Change Average (Rasch)	11.9	12.4	12.1	12.0	12.1
Case Count	26,472	26,654	26,927	25,620	21,629
Number of Facilites at or above Expectation	72	80	77	74	83
Number of Facilities below Expectation	56	61	70	69	71
Percent of Facilities at or above Expectation	56.3%	56.7%	52.4%	51.7%	53.9%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the motor measure for the years 2010 to 2014.

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2010		2011		2012		2013		2014	
	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)
Gender										
Male	7,668	9.0	7,705	12.3	7,617	12.0	6,489	11.9	5,100	12.2
Female	13,768	8.5	13,730	12.5	13,061	12.4	10,362	12.4	8,204	12.2
Ethnicity										
White	14,461	12.0	14,422	12.1	13,586	12.1	9,766	12.0	8,014	11.6
Black	2,073	12.6	2,273	13.9	1,997	13.3	1,609	13.2	1,453	13.2
Hispanic	370	14.3	400	13.9	353	14.1	216	12.2	140	12.5
Other Ethnicity	9,568	11.6	9,559	12.4	10,991	11.9	14,029	12.0	12,022	12.2
Payer Source										
Medicare	18,658	12.1	19,261	12.5	19,898	12.2	18,842	12.2	15,577	12.2
Medicaid	669	9.5	525	13.0	566	13.9	519	13.8	514	14.0
Commercial	1,826	12.0	2,032	12.6	2,052	12.4	2,247	11.8	1,799	11.9
Blue Cross	1,168	14.4	845	14.4	876	14.2	999	13.3	526	12.6
Other Payer	4,151	11.0	3,991	11.3	3,535	10.8	3,013	10.5	3,213	10.9
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	3,481	11.0	3,310	10.3	3,784	10.2	3,539	9.8	3,437	10.0
P02 (NY, NJ, PR)	9,099	13.5	7,581	13.3	6,031	13.7	6,290	13.4	4,426	12.8
P03 (PA, WV, VA, DE, MD, DC)	1,793	11.3	1,489	11.6	1,565	13.1	1,721	12.3	1,198	12.7
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	8,057	10.7	7,542	13.6	7,401	12.3	8,759	11.8	7,405	12.3
P05 (MN, WI, IL, IN, MI, OH)	3,728	12.3	3,290	11.8	3,313	12.3	4,289	12.3	4,907	12.4
P06 (NM, OK, AR, LA, TX)	29	12.1	2,015	10.9	2,685	11.4	383	12.3	0	-
P07 (NE, IA, KS, MO)	285	8.8	1,381	9.7	2,124	10.4	639	11.4	135	9.8
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	33	12.2
P09 (CA, NV, AZ, HI)	0	-	46	16.6	24	16.0	0	-	88	10.9
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the motor measure is a new measure and has not been used as a stand-alone tool. However, all of the items within the measure are included in a larger instrument (the FIM instrument) which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the motor measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and use of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Skilled Nursing Facilities:

1. Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.
2. Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.
3. Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.
4. Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

5. Munin MC, Begley A, Skidmore ER, Lenze EJ. Influence of rehabilitation site on hip fracture recovery in community-dwelling subjects at 6-month follow-up. *Archives of physical medicine and rehabilitation.* Jul 2006;87(7):1004-1006.
6. Munin MC, Seligman K, Dew MA, et al. Effect of rehabilitation site on functional recovery after hip fracture. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):367-372.
7. Nelson DL, Melville LL, Wilkerson JD, Magness RA, Grech JL, Rosenberg JA. Interrater reliability, concurrent validity, responsiveness, and predictive validity of the Melville-Nelson Self-Care Assessment. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association.* Jan-Feb 2002;56(1):51-59.
8. Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Archives of physical medicine and rehabilitation.* Oct 1996;77(10):1056-1061.
9. Vincent KR, Vincent HK. A multicenter examination of the Center for Medicare Services eligibility criteria in total-joint arthroplasty. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Jul 2008;87(7):573-584.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Barnes C, Conner D, Legault L, Reznickova N, Harrison-Felix C. Rehabilitation outcomes in cognitively impaired patients admitted to skilled nursing facilities from the community. *Archives of physical medicine and rehabilitation.* Oct 2004;85(10):1602-1607.

OBJECTIVE: To examine the outcomes of patients with varying levels of cognitive impairment who received rehabilitation in skilled nursing facilities (SNFs). DESIGN: A retrospective analysis of the records of people admitted to SNFs for rehabilitation. SETTING: Seven SNFs in Colorado. PARTICIPANTS: Community-dwelling persons (N=7159), 65 years of age and older, admitted for rehabilitation after a hospitalization or decline in function between May 1998 and May 2002. Interventions Not applicable. MAIN OUTCOME MEASURES: Cognitive impairment was assessed using a 4-level categorization of the FIM instrument cognitive score at admission. Functional gain was measured using the FIM. Community discharge was measured as the proportion of patients discharged to home, board and care, or assisted living facility. Rehabilitation progress was measured as the number of FIM points gained per day. RESULTS: Significant functional gains were made during rehabilitation in motor and cognitive FIM scores, regardless of cognitive impairment. The most cognitively impaired patients required more rehabilitation intervention, achieved less FIM gain, and were less likely to be discharged to the community. The strongest predictors of FIM gain were the amount of therapy hours and admission cognitive FIM score. The strongest predictors of discharge to the community were the discharge total FIM score and age. The strongest predictors of adequate rehabilitation progress were medical complexity and admission cognitive FIM score. CONCLUSIONS: Patients with cognitive impairment were able to recover function with rehabilitation intervention. Patients with a more serious cognitive impairment received more rehabilitation intervention than patients with less impairment. Outcomes were predicted by admission and rehabilitation measures that were qualitatively different from other discharge outcomes. Health care professionals need to consider these factors as they create a rehabilitation plan of care for patients with cognitive impairment.

Chen CC, Heinemann AW, Granger CV, Linn RT. Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Archives of physical medicine and rehabilitation.* Nov 2002;83(11):1514-1523.

OBJECTIVES: To document patient, program characteristics, and therapy service provision in subacute rehabilitation across 3 types of facilities that provide subacute rehabilitation, to examine the determinants of therapy intensity, and to evaluate the

contribution of rehabilitation services to functional gains. DESIGN: A retrospective study linking administrative billing data and patients' functional assessment records. SETTING: Twenty facilities part of the Uniform Data System for Medical Rehabilitation (UDSMR) subacute database PARTICIPANTS: A total of 1976 billing records of patients with stroke, orthopedic, and debility impairments, discharged in 1996 and 1997, were retrieved and linked with the FIM trade mark instrument ratings from UDSMR subacute database. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: Total therapy intensity and Rasch-transformed FIM domain gains (ie, gains in self-care, mobility, cognition). RESULTS: Therapy intensity was mostly determined by impairment and facility type, although variances explained by the predictors were small. Patients in all 3 impairment groups made functional gains; gains were related weakly, although significantly, to therapy intensity and rehabilitation duration after controlling for other variables. CONCLUSIONS: The provision of rehabilitation therapies varied across facilities. Skilled nursing facilities with subacute rehabilitation units tended to provide more therapies than subacute units in acute or rehabilitation hospitals.

Jette DU, Warren RL, Wirtalla C. The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Archives of physical medicine and rehabilitation.* Mar 2005;86(3):373-379.

OBJECTIVE: To examine the relation between therapy intensity, including physical therapy (PT), occupational therapy (OT), and speech and language therapy (SLT), provided in a skilled nursing facility (SNF) setting and patients' outcomes as measured by length of stay (LOS) and stage of functional independence as measured by the FIM instrument. DESIGN: A retrospective analysis of secondary data from an administrative dataset compiled and owned by SeniorMetrix Inc. SETTING: Seventy SNFs under contract with SeniorMetrix health plan clients. PARTICIPANTS: Patients with stroke, orthopedic conditions, and cardiovascular and pulmonary conditions (N=4988) covered by Medicare+Choice plans, and admitted to an SNF in 2002. INTERVENTIONS: Not applicable. MAIN OUTCOMES MEASURES: LOS and improvement in stage of independence in the mobility, activities of daily living (ADLs), and executive control domains of function as determined by the FIM instrument. RESULTS: Higher therapy intensity was associated with shorter LOS (P <.05). Higher PT and OT intensities were associated with greater odds of improving by at least 1 stage in mobility and ADL functional independence across each condition (P <.05). The OT intensity was associated with an improved executive control stage for patients with stroke, and PT and OT intensities were associated with improved executive control stage for patients with cardiovascular and pulmonary conditions (P <.05). The SLT intensity was associated with improved motor and executive control functional stages for patients with stroke (P <.05). Therapy intensities accounted for small proportions of model variances in all outcomes. CONCLUSIONS: Higher therapy intensity was associated with better outcomes as they relate to LOS and functional improvement for patients who have stroke, orthopedic conditions, and cardiovascular and pulmonary conditions and are receiving rehabilitation in the SNF setting.

Latham NK, Jette DU, Warren RL, Wirtalla C. Pattern of functional change during rehabilitation of patients with hip fracture. *Archives of physical medicine and rehabilitation.* Jan 2006;87(1):111-116.

OBJECTIVE: To examine the rate of functional change in 2 domains, activities of daily living (ADLs) and mobility, over 2 time periods during hip fracture rehabilitation. DESIGN: Retrospective analysis of data contained in an administrative dataset. SETTING: Seventy skilled nursing facilities (SNFs). PARTICIPANTS: People (N=351) receiving rehabilitation in SNFs from March 1998 to February 2003 after hip fractures. INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURE: Rate of change in scores in the ADL and mobility domains of the FIM instrument during 2 time intervals of rehabilitation. RESULTS: The rate of functional change across 2 time intervals was constant for mobility (mean change in FIM points per day, .46 vs .49), but declined in the second time period for ADLs (mean change in FIM points per day, .55 vs .41). Executive function, length of stay (LOS), and medical complexity were related to rate of change in mobility, and baseline ADLs, executive function, living setting, and LOS were related to rate of change in ADLs. There was an interaction between rehabilitation phase and baseline mobility. People with lower baseline mobility had an increased rate of change during the second interval (mean change in FIM points per day, .41 vs .55), whereas those with higher baseline mobility had a decreased rate of change (mean change in FIM points per day, .50 vs .43). CONCLUSIONS: The pattern of functional change over time differed for ADL and mobility domains, and for specific groups of patients. The results have implications for goal setting and discharge planning.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2775
De.2. **Measure Title:** Functional Change: Change in Motor Score for Skilled Nursing Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
De.3. **Brief Description of Measure:** Change in rasch derived values of motor function from admission to discharge among adult short term rehabilitation skilled nursing facility patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 12 items:Feeding, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
1b.1. **Developer Rationale:** The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing functional items presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the measure. Currently more than 150 SNFs are utilizing the items in our proposed measure for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

S.4. **Numerator Statement:** Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the SNF or patients who died within the SNF are excluded.
S.7. **Denominator Statement:** Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.
S.10. **Denominator Exclusions:** Patients age at admission less than 18 years old
Patients who died in the SNF.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Motor_SNF-635950325390738085.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The current mandated quality measures for Skilled Nursing Facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing functional items presently used across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the measure. Currently more than 150 SNFs are utilizing the items in our proposed measure for outcomes purposes; therefore, it should not be difficult for all SNFs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.

We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

While this is a new measure, UDSMR has historical data on all 12 items, and we are able to give information on the measure. See measure evaluation form for the trending data.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

See the measure evaluation sheet for disparity data overtime for the measure.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_Submission-635749892715380581.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the SNF or patients who died within the SNF are excluded.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in the SNF. The numerator is the average change in rasch derived motor functional score from admission to discharge for each patient at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted expected change in rasch derived values, adjusted for SNF-CMG (Skilled Nursing Facility Case Mix Group), based on impairment type, admission functional status, and age.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all short term rehabilitation patients at the skilled nursing facility, at least 18 years old, who did not die in the SNF. Impairment type is defined as the primary medical reason for the SNF short term rehabilitation stay (such as stroke, joint replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum 12 items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) at the facility level. Age is the age of the patient at the time of admission to the SNF. The denominator is meant to reflect the expected motor functional change score at the facility, if the facility had the same distribution of SNF-CMGs (based on impairment type, functional status at admission, and age at admission). This adjustment procedure is an indirect standardization procedure (observed facility average/expected facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients age at admission less than 18 years old
Patients who died in the SNF.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through the MDS.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the SNF-CMGs in the excel file provided.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standarization procedure (observed facility average/expected facility average). The numerator is the facility´s average motor functional change score. The denominator is meant to reflect the expected motor functional change score at the facility, if the facility had the same distribution of SNF-CMGs (impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all short term rehabilitation patients during the assessment time frame (12 months).
2. Exclude any patients who died in the SNF.
3. Exclude any patients who are less than 18 at the time of admission to the SNF.
3. Calculate the total motor change score for each of the remaining patients (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in the attached excel file).
5. Calculate the average rasch derived motor change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average motor change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average motor change score/facility´s national expected motor change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the appendix.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Motor_Total_SNF-635950325606769000.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Motor Score in Skilled Nursing Facilities
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation

counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® ("FIM") instrument data from inpatient rehabilitation facilities, long term acute care facilities, and skilled nursing facilities from the Uniform Data System for Medical Rehabilitation. The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the motor measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2002-2013 were used in examining the data trends over time using the motor measure and patient outcomes of inpatient rehabilitation

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice

☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☒ other: patient level/aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term acute care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument (the tool used for this measure) is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM® instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 12 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for the inpatient rehabilitation setting. However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our motor measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

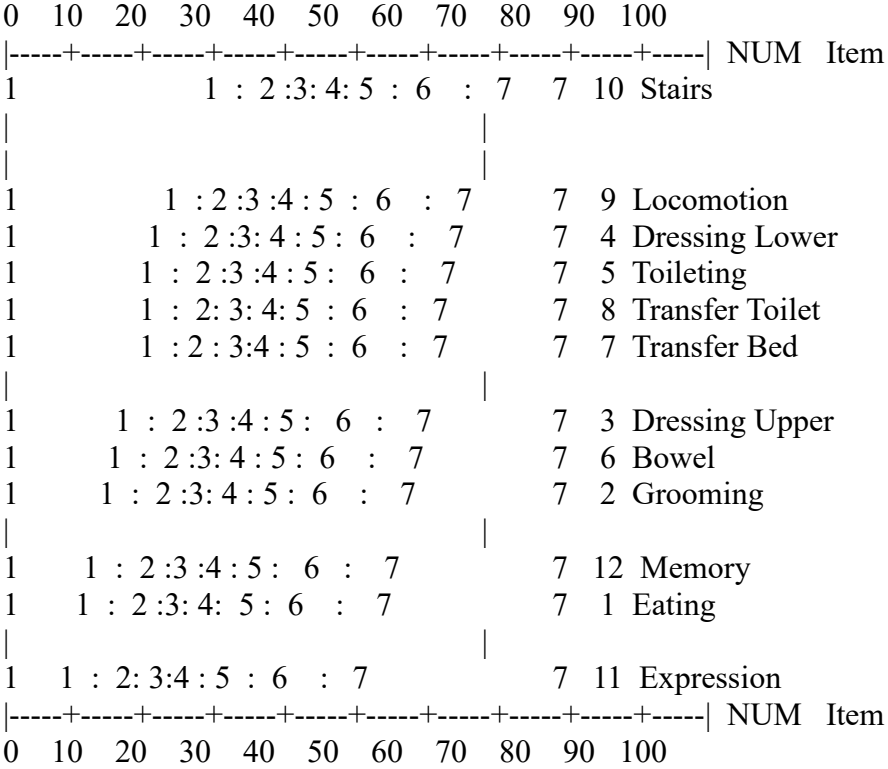
Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (12 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory and Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistantcompared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.94. The Cronbach Alpha reliability statistic was 0.95. Item correlations within the measure ranged from 0.65 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure, we set the Rasch scale at 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.



The ruler shows that the easiest item is Expression, and the hardest Stairs and that the distances between a level 1 and 2 and 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 12 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
12	.00	17.24	37	37.90	2.28	62	52.00	2.63
13	10.58	8.94	38	38.43	2.27	63	52.73	2.67
14	16.04	6.04	39	38.96	2.26	64	53.47	2.72
15	19.04	4.85	40	39.48	2.25	65	54.25	2.76
16	21.12	4.19	41	40.00	2.24	66	55.05	2.81
17	22.75	3.78	42	40.51	2.23	67	55.88	2.86
18	24.11	3.49	43	41.03	2.23	68	56.74	2.92
19	25.29	3.28	44	41.54	2.23	69	57.64	2.99
20	26.34	3.12	45	42.05	2.23	70	58.58	3.06
21	27.31	2.99	46	42.57	2.23	71	59.57	3.15
22	28.20	2.89	47	43.08	2.24	72	60.63	3.25
23	29.03	2.80	48	43.60	2.25	73	61.76	3.37
24	29.82	2.73	49	44.13	2.26	74	62.98	3.50
25	30.57	2.66	50	44.66	2.28	75	64.30	3.66
26	31.28	2.61	51	45.20	2.29	76	65.75	3.85
27	31.97	2.56	52	45.74	2.31	77	67.37	4.07
28	32.63	2.51	53	46.30	2.34	78	69.19	4.34
29	33.28	2.48	54	46.87	2.36	79	71.29	4.69
30	33.90	2.44	55	47.45	2.39	80	73.79	5.17
31	34.51	2.41	56	48.05	2.42	81	76.91	5.87
32	35.10	2.38	57	48.66	2.45	82	81.15	7.07
33	35.68	2.36	58	49.29	2.49	83	88.16	9.82
34	36.25	2.34	59	49.94	2.52	84	100.00E	17.75
35	36.81	2.32	60	50.61	2.56			
36	37.36	2.30	61	51.29	2.60			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As indicated previously, the reliability of the FIM instrument is well known. The results of the analysis for the measure proposed show the reliability holds even when looking at a subset of FIM instrument items.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the motor measure with the FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new motor measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (motor items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the motor measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM® instrument (the gold standard)), and likelihood of discharge to the community setting Linear regression was used to determine functional change, whereas the change in motor was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission motor total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the motor measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was highly correlated with the FIM total, 0.932 ($p < 0.001$) and 0.952 ($p < 0.001$), respectively. The correlations remained highly significantly within each venue of care; IRFs, 0.927 ($p < 0.001$) and 0.963 ($p < 0.001$); LTACs, 0.935 ($p < 0.001$) and 0.953 ($p < 0.001$); SNFs, .944 ($p < 0.001$) and .947 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM totals, the r-square values were extremely high, 0.962 and 0.982, respectively. The values remained high at the venue specific level as well; IRFs, 0.945 and 0.974; LTACs, 0.968 and 0.985; SNFs, 0.960 and 0.980.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was very high, 0.866 ($p < 0.001$). In addition, by venue, the correlations remained strong; IRFs, 0.868 ($p < 0.001$); LTACs, 0.887 ($p < 0.001$); SNFs, 0.837 ($p < 0.001$). The linear regression showed high r-squared values as well; all venues, 0.751; IRFs, 0.754; LTACs, 0.786; SNFs, 0.701.

Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in our measure has good predictive ability for discharge setting (community), with a C-statistic of 0.77. By venue, the results are similar; IRFs, 0.75; LTACs, 0.754.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results show good validity across all analyses. The r-squared values were all above 0.8, meaning that the percent of variance explained in the dependent variables by our measure were all more than 80%. In addition, the predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients that died in the post-acute care setting (an unanticipated outcome) and patient aged 18 years and older, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 1 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Skilled Nursing Facility Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national SNF-CMG-specific values by facility-specific SNF-CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The skilled nursing facility case-mix group (SNF-CMG) classification system groups similarly impaired patients based on functional status at admission or patient severity. Patients within the same SNF-CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a SNF-CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM instrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all CMGs.)

See file uploaded in S.15 for calculations.

The SNF-CMGs are groupings specific to skilled nursing facilities, although they are similar and easily comparable to the CMGs used in inpatient rehabilitation facilities.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.

2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.
4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Over 150 SNFs currently collect data on the items in our proposed measure for quality benchmarking, both internally and as a national benchmarking system. Therefore, the feasibility of this measure is sound.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The Functional Change: Change in Motor Score form (this form includes the items for the motor measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) UDSMR http://www.udsmr.org/ Quality Improvement (Internal to the specific organization) UDSMR www.udsmr.org

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for SNFs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for skilled nursing facilities, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

The items of our proposed measure are part of the FIM system, which is in use in nearly 150 SNFs in the United States. Outcomes based on the items are currently used for Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and Quality Improvement (Internal to the specific organization) for those SNFs utilizing the FIM system.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
This is a new measure.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences to individuals or populations during the testing of this measure as previously collected data was used.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Functional_Change_Appendix-635749870363739883.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 03, 2016
Ad.4 **What is your frequency for review/update of this measure?** unknown, new measure
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

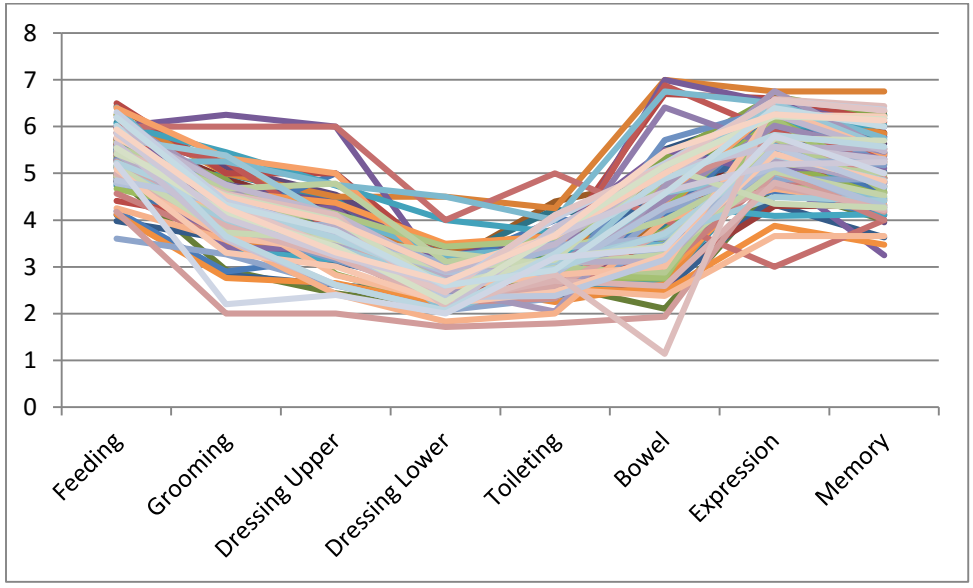
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

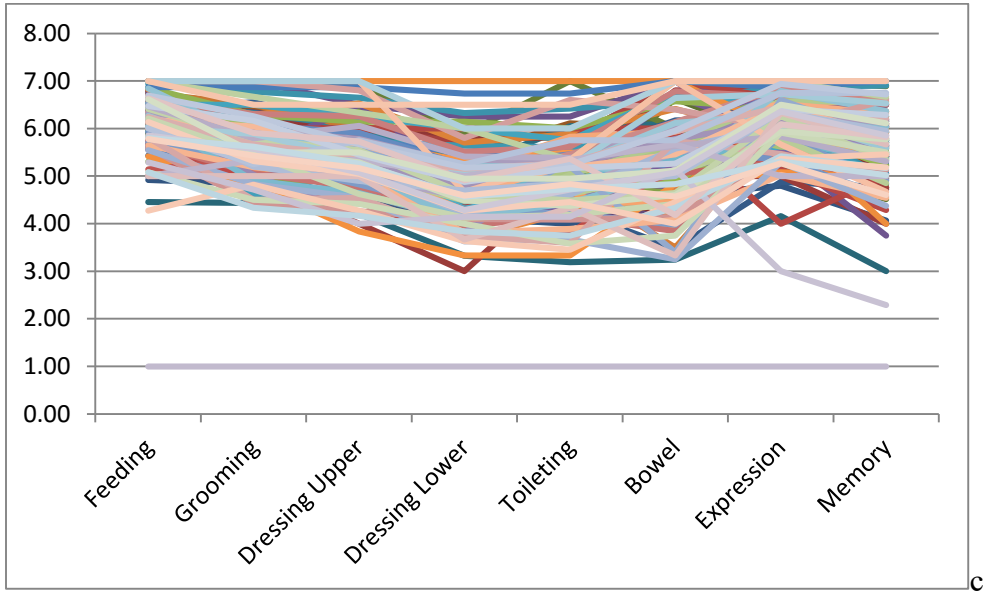
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Skilled Nursing Facilities, Change in Function: Mobility Measure for Skilled Nursing Facilities and the Change in Function: Motor Measure for Skilled Nursing Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

Self-Care Graph: Admission (Year 2015)

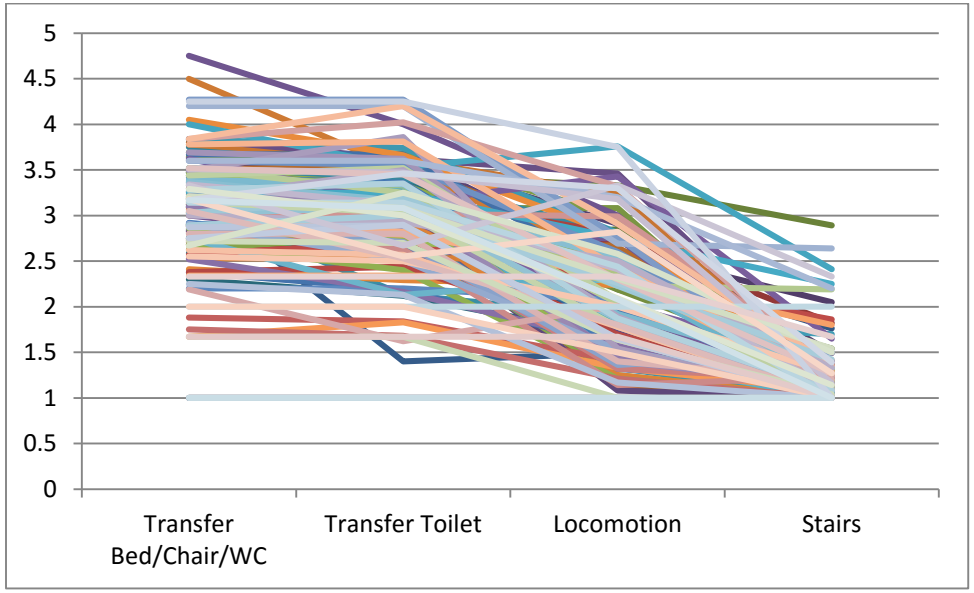


Self-Care Graph Discharge (Year 2015)

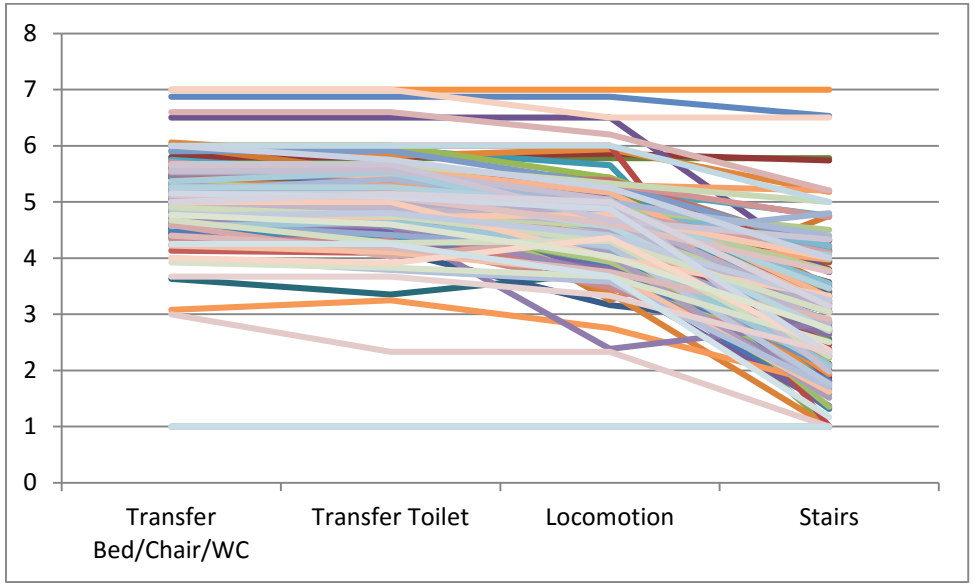


c

Mobility Graph: Admission (Year 2015)



Mobility Graph: Discharge (Year 2015)



Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the SNF venue of care.

Self-Care Mean Fit Statistics

```
TABLE 3.1 Self Care 8 Items                    ZOU018WS.TXT  Mar 19  9:16 2015
INPUT: 3096 Person  8 Item  REPORTED: 3094 Person  8 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2969 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      36.6       8.0       50.76     3.96       .96    -.1   1.02     .0 |
| S.D.      11.5        .3       13.60     1.46       .71    1.2    .82    1.2 |
| MAX.      55.0       8.0       87.04    10.90      6.32    5.4   8.33    6.2 |
| MIN.       8.0       3.0       11.87     3.00       .05   -3.9    .05   -3.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.60 TRUE SD   12.80  SEPARATION  2.78  Person RELIABILITY  .89 |
|MODEL RMSE   4.22 TRUE SD   12.93  SEPARATION  3.06  Person RELIABILITY  .90 |
| S.E. OF Person MEAN = .25                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     50 Person
  MINIMUM EXTREME SCORE:     75 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      36.2       8.0       50.33     4.59                               |
| S.D.      12.4        .3       16.71     3.40                               |
| MAX.      56.0       8.0      100.06    19.89                               |
| MIN.       8.0       3.0        -.06     3.00       .05   -3.9    .05   -3.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.99 TRUE SD   15.60  SEPARATION  2.61  Person RELIABILITY  .87 |
|MODEL RMSE   5.71 TRUE SD   15.70  SEPARATION  2.75  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .30                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93
```

Mobility Mean Fit Statistics

```
                                                  03-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       13.8       3.7       31.44     4.51       .94    -.3    .94    -.2 |
| S.D.        6.2        .5       16.49     1.26      1.27    1.4   1.34    1.2 |
| MAX.       27.0       4.0       87.88     9.51      9.90    5.8   9.90    8.5 |
| MIN.        2.0       1.0        8.08     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89 |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       12.2       3.7       26.70     5.88                              |
| S.D.        6.9        .6       19.75     3.22                              |
| MAX.       28.0       4.0       99.95    13.79                              |
| MIN.        1.0       1.0         .02     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87 |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

TABLE 3.1 All Facilities 12 items ZOU439WS.TXT Mar 19 9:43 2015
INPUT: 3096 Person 12 Item REPORTED: 3094 Person 12 Item 7 CATS WINSTEPS 3.73

SUMMARY OF 3013 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	49.2	11.6	45.63	2.83	.99	-.1	1.06	.0
S.D.	17.6	.7	12.31	.98	.67	1.4	.91	1.4
MAX.	83.0	12.0	88.22	9.85	5.13	5.2	9.90	7.7
MIN.	10.0	4.0	10.53	2.23	.09	-4.2	.11	-3.8
REAL RMSE	3.30	TRUE SD	11.86	SEPARATION	3.59	Person RELIABILITY	.93	
MODEL RMSE	2.99	TRUE SD	11.94	SEPARATION	3.99	Person RELIABILITY	.94	
S.E. OF Person MEAN = .22								

MAXIMUM EXTREME SCORE: 7 Person
MINIMUM EXTREME SCORE: 74 Person
LACKING RESPONSES: 2 Person

SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	48.4	11.7	44.66	3.21				
S.D.	18.3	.7	14.26	2.51				
MAX.	84.0	12.0	100.06	17.81				
MIN.	10.0	4.0	-.05	2.23	.09	-4.2	.11	-3.8
REAL RMSE	4.30	TRUE SD	13.59	SEPARATION	3.16	Person RELIABILITY	.91	
MODEL RMSE	4.07	TRUE SD	13.66	SEPARATION	3.36	Person RELIABILITY	.92	
S.E. OF Person MEAN = .26								

Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2776
Measure Title: Functional Change: Change in Motor Score in Long Term Acute Care Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Brief Description of Measure: Change in rasch derived values of motor function from admission to discharge among adult long term acute care facility patients aged 18 years and older who were discharged alive. The timeframe for the measure is 12 months. The measure includes the following 12 items: Feeding, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
Developer Rationale: The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or a less intensive setting of care. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing items which are presently collected across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. There are LTACs that are currently collecting data on the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting or less intensive setting upon discharge.

Numerator Statement: Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the LTAC or patients who died within the LTAC are excluded.
Denominator Statement: Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age.
Denominator Exclusions: Patients age at admission less than 18 years old
Patients who died in the LTAC.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer states "The primary aim of rehabilitation at LTACs is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or LTAC stay."
- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. While the FIM tool is presently primarily used in inpatient rehabilitation facilities, they state there are LTACs collecting data using the FIM. They provide a list of 3 peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in LTACs.
- The items that comprise the motor measure are as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

According to the developer, "The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers."

While this is a new measure, UDSMR has been collecting data on the FIM for more than 20 years so they have historical data to report. The most recent data reported is from 2011 and indicates more than 60% of cases are below expectation. They offer the following table for LTAC patients:

Year	2007	2008	2009	2010	2011
Motor Change Average (Rasch)	11.2	11.5	12.0	11.3	12.1
Case Count	5,807	5,303	4,996	4,861	4,598
Number of Facilites at or above Expectation	9	8	7	7	5
Number of Facilities below Expectation	9	8	9	7	8
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	50.0%	38.5%

The developer provided additional documentation stating that the mean score is 49.2, the standard deviation is 17.6, the max is 83.0 and the minimum is 10.0.

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the motor

score measure for the years 2010 to 2014. However, information is not provided on whether the differences are statistically significant.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. Evidence to Support Measure Focus

Comments:

**Relationship between measured outcome (change in motor function score from admission to discharge) and LTAC care is identified, although not specifically. LTAC services are generally referred to, but it is assumed that this measure is meant to demonstrate relationship between provision of OT and PT services within an LTAC setting as intervention with impact on overall self care and mobility elements of motor score. Evidence referenced supports relationship.

Measure calculated from inventory score at admission and at discharge. Developer references interim use to assess progress, but no evaluation of such data. Why? If this is a 12-month measure, wouldn't periodic review (even of non discharged patients) be valuable to assessment of quality care by the LTAC?

How are discharges to hospitalization handled in this measure calculation (I assume they aren't counted, but it's not addressed)

Cmte question: Is there at least one thing the provider can do to achieve change in the measure? Presumably, provision of high quality OT, PT and similar interventions can lead to improvement in motor function and self care (less clear that cognition/memory can be influenced).

Journal articles referenced demonstrate the value of the calculation of a score on predictive value of discharge readiness and over improvement in assessment criteria over LOS. None of the articles addresses how such measurement improves assessment of quality of care provided OR comparability of service provision from one facility vs. another.

**Given Long Term Acute Care facilitates' emphasis on "restorative or maintenance of function affected by the patients' illness or injury," the measure's focus is of great importance.
The developers provide a flowchart that links rehabilitation therapy to measured outcomes (change in motor functioning), which rationalizes use of the measure.

1b. Performance Gap

Comments:

**The developer notes and provides graphs that the patterns of motor score change vary between facilities but tend to follow the same pattern between admission and discharge. I'm not sure what can be said about using such data to compare quality both within a facility (ie between CMGs or o see patterns of all cases to identify outliers and pinpoint reasons for those different outcomes) and comparing across facilities. The developers reference that this measure has highest value for internal facility level evaluation and provision of substantiation of value of their interventions to payers (including CMS). There is a gap in care that warrants a national performance measure. The developers call it ability to measure adequacy of rehabilitation services and functional status of LTAC patients. In addition, there is a gap as to what mix, duration and intensity of interventions have the greatest impact on function (motor, self care, cognition) and can provide measurable changes that lead to shorter LOS, greater rates of discharge to community.

**Developers provide annual data on the percentage of facilities at or above expectation, which indicates that there is substantial room for improvement on the PM scores.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, OASIS

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:
  - Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients, over Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age.
  - Average is calculated as (sum of change at the patient level/total number of patients).
- The calculation algorithm is included.
- Patients under age 18 and patients who died in the LTAC are excluded.
- A data dictionary is included.
- The measure is stratified by risk category using an indirect standardization procedure (observed facility average/expected facility average)

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented using inter and intra-rater reliability
- This measure uses a subset of the FIM, so a Rasch analysis was conducted to test:
  - the psychometric properties of the subset of 12 items within the three venues of post-acute care, IRFs, LTACs, and SNFs
  - the measure reliability at both the person and item level
  - to determine the fit of each item within the measure (12 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory and Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations.
- Internal consistency demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but is working to do so prior to the in-person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities*

(perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities. UDSMR has indicated they are working to fulfill these data needs.

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.94. The Cronbach Alpha reliability statistic was 0.95. Item correlations within the measure ranged from 0.65 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empirical testing of data elements (box 2) -> TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Developers used concurrent validity of the FIM total score (all 18 items) with the FIM motor score: the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (motor items).
- Predictive validity of the motor score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:

- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for LTACs are 0.935 ($p < 0.001$) at admission and 0.953 ($p < 0.001$) at discharge. For predicative validity, LTACs scored 0.887 ($p < 0.001$).
- The r-squared values were all above 0.8, meaning that the percent of variance explained in the dependent were all more than 80%. For LTACs, the r-squared values at admission were 0.968 and at discharge 0.985 for functional gain. The C-statistic for LTACs is 0.754.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the facility were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developer states the following risk adjustment method: To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The case mix group classification system groups similarly impaired patients based on functional status at admission or patient severity. This is used for SNFs and IRFs, and the same procedure will be applied to the LTACs. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *No information is provided on risk adjustment for SDS factors. Do you think the measure should include SDS factors in the risk adjustment? Why or why not?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provided additional information in an addendum, including "graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers".

Question for the Committee:
Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**data element definitions: clear
codes – clear
calculation steps – clearly outlined
risk/case mix adjustment – complex but definition seems clear
no sampling required as all cases within a CMG are included unless (<18 yo or patient died while in LTAC).
Concern about consistent implementation: Developer references common process for LTAC to assess these functional areas but it's not clear that this is universal OR that a common tool exists to do this as this measure relies on a proprietary subscriber service provided instrument and supporting analysis service.
**There is no explicit discussion of the target population's input into specifications or the identification of meaningful cut-points or change in FIM scores.

2a2. Reliability Testing

Comments:

**Empirical testing of data was not provided by the developer and is pending per NQF

Examples of Rasch analyses comparing PAC, IRF, LTAC and SNF sites were clear and sample size seemed consistent and reasonable. Noting that data came from subscribers of the developer so it's not clear if that would affect generalizability to LTACs.
The worksheet referenced that reliability tested at the person and item level, but this is indicated as a facility level measure. Shouldn't testing address facility level comparison?

**The developers cite the extensive literature on the reliability of the full FIM. This seems of little value since the proposed measure uses a subset of FIM items. That said, they provide adequate evidence of measurement reliability at the individual level for the 12 FIM items used in this measure. They describe results of a rating pattern analysis (facilities average scores for each item) as evidence of reliability at the organization level. I look forward to learning more about this procedure.

2b.2 Validity Testing
Comments:
**Specifications appear consistent with evidence provided.
Testing found measure score had predictive value for functional change outcomes and likelihood of discharge, both goals of rehabilitation services.
Change in score can be seen as indication of quality services, but do other factors need to be assessed for their impact on functional status? Value of family involvement and reinforcement, patient engagement in goal setting and therapy provided, for example.

**Item face validity is strong.

Developers assessed concurrent validity by evaluating associations between scores on the motor measure and the full FIM. Likewise, they evaluate predictive validity by comparing gains on the proposed measure to overall FIM gains. In my opinion, these analyses do not provide compelling evidence of measurement validity since the motor measure is comprised of a subset of FIM items (of course they are highly correlated).
The measure's prediction of discharge disposition is more compelling.

No evidence of validity at the organization level is provided.

2b3.-2b7. Test Related to Potential Threats
Comments:
**2b3-7 Threats to Validity
Exclusions are consistent and do not appear to be inappropriate

Question about how discharges for hospitalization are handled within this measure or are those de facto exclusion that should be identified.

2b4 risk adjustment strategy is defined and it appears that such criteria would be present at the start of care.

No information on risk adjustment for SDS factors. This is lacking in the developer's package and should be addressed despite the developer's observations that existing motor measure data doesn't indicate disparities in average score based on gender or ethnicity.

2b5 Difference

The measure as defined would identify differences in motor score over time within case mix groups in an LTAC. It would demonstrate change in functional capability and readiness for discharge. Its hard to say this measure would show meaningful differences in quality because the developer didn't' show us data showing one facilities' data by

CMG for example (they showed by gender, ethnicity, payer source and CMS region. The developer says (p 12) that the data provided demonstrates change at the case level rather than facility level outcomes and comparisons. They suggest that this can be provided and such data should be requested to facilitate the committee's analysis on this point.

**The developers note that all variables are required, which is presumed to reduce missing data. Case-deletion is used for patients with missing data. The developers failed to present data on the percentage of cases that are excluded due to missingness. Therefore, the degree to which missing data biases PM scores is unclear.

The case-mixed adjustment procedures are not clearly rationalized. It is only noted that the apply a "standard procedure."

There is no information about what constitutes a meaningful change in PM scores and no description of the measure's ability to detect statistically significant or clinically/practically meaningful differences in PM scores across organizations or within an organization over time.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- The developer reports there are LTACs currently using the FIM
- Commercial use requires a license agreement and has a fee. The developer reports the following:
    - The Functional Change: Change in Motor Score form (this form includes the items for the motor measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**Current use of FIM by LTACs (is this universal ?) indicates data elements in calculating the score are generated and used in care delivery. There is no commentary on where gaps might exist (ie do ALL LTACs assess this within 24 hours of admission?) and whether such data are universally accessible by EHR. Again, the reference is to LTACs that are subscribers to developers' services currently.

There is a license and fee associated with commercial use of this tool/measure and the expectation would be that a copyrighted assessment form be used as is. The developer does not address possible differences in clinical practice at non subscriber LTACs and this is an important area of inquiry for the committee to discern whether this would be readily implemented in all such settings.

The developer indicates this measure is currently used for internal reporting and benchmarking purposes by subscribers. That purpose is different than quality assessment and comparison across facilities.

This measure requires complex calculation of Rasch score and complex risk adjustment algorithm. It's unclear that this would be readily implemented in average LTAC without the express services (by subscription) of the developer. Concern is for conflict of interest as the developer can gain significantly from establishment of this measure as a national standard of quality.

**The FIM is already required, which substantially enhances the feasibility of this measure. The FIM is a proprietary instrument, but its use in this context appears to be permissible at low or no cost.
Calculation of facility's adjusted expected change in Rasch derived values does not seem straight forward. How will facilities be trained in these procedures?

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

- The measure is currently used for internal reporting and national benchmarking by LTACs who subscribe to the UDSMR software/outcomes reporting.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results

- New measure – not available. While a new measure to NQF, the developer does provide trending data for the rasch derived scores from 2007-2011:

Year	2007	2008	2009	2010	2011
Motor Change Average (Rasch)	11.2	11.5	12.0	11.3	12.1
Case Count	5,807	5,303	4,996	4,861	4,598
Number of Facilites at or above Expectation	9	8	7	7	5
Number of Facilities below Expectation	9	8	9	7	8
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	50.0%	38.5%

Unexpected findings (positive or negative) during implementation

- None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4 Usability and Use

Comments:

**Developer indicates no plan for public reporting using this measure (it's not currently used in this way) and it's unclear what public reporting value the scores would have (they are hard to understand and the graphs the developer provided do not demonstrate how they would be used by public to understand value of rehabilitation services in an LTAC setting.

Criterion 5: Related and Competing Measures

Related or competing measures

None reported

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Motor Score for Long Term Acute Care Facilities

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*

- *A separate evidence form is required for each component measure unless several components were studied together.*
- *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*

- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

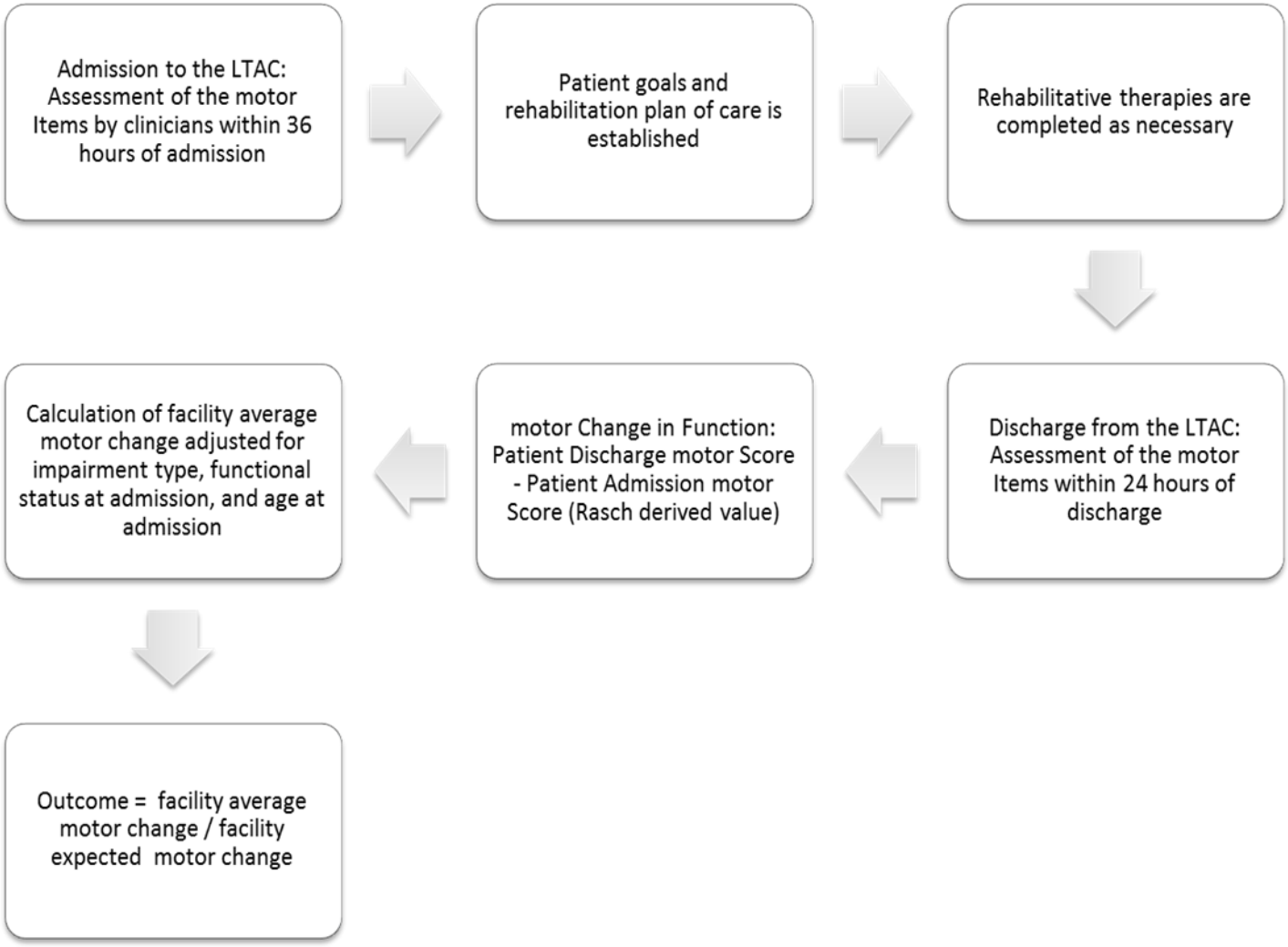
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Long Term Acute Care Hospitals (LTACs) are one part of a multi-level post-acute care continuum. The primary aim of rehabilitation at LTACs is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or LTAC stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are LTACs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for patients seen at LTACs. To date, the motor measure has not been reported on as a stand-alone measure. However, the items of the motor measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument. The motor measure is intended to be administered within 24 hours of the patient's admission to the LTAC and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the motor measure is as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.). All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for LTAC patients:



UDSMR has been a data repository for the FIM instrument among LTAC patients, of which the items of the motor measure are nested within for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the self-care measure for the years 2007 to 2011 for LTAC patients:

Year	2007	2008	2009	2010	2011
Motor Change Average (Rasch)	11.2	11.5	12.0	11.3	12.1
Case Count	5,807	5,303	4,996	4,861	4,598
Number of Facilites at or above Expectation	9	8	7	7	5
Number of Facilities below Expectation	9	8	9	7	8
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	50.0%	38.5%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the motor measure for the years 2007 to 2011:

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2007		2008		2009		2010		2011	
	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)	Case Count	Motor Change Average (Rasch)
Gender										
Male	3,126	11.6	2,897	11.9	2,724	12.5	2,641	11.8	2,493	12.3
Female	2,676	10.7	2,398	11.2	2,267	11.5	2,215	10.7	2,101	11.7
Ethnicity										
White	4,653	11.2	4,346	11.5	3,895	12.1	3,606	11.1	3,508	12.0
Black	636	11.4	547	12.0	538	12.3	463	11.2	379	12.2
Hispanic	62	10.6	61	13.4	56	12.8	81	12.0	47	12.3
Other Ethnicity	456	10.4	349	11.3	507	11.3	711	12.0	664	12.2
Payer Source										
Medicare	3,444	9.8	3,075	10.0	2,264	10.3	2,222	10.0	2,342	10.7
Medicaid	366	13.1	337	13.1	321	13.7	246	13.1	225	14.0
Commercial	679	12.7	641	13.7	657	12.6	631	12.3	535	14.2
Blue Cross	588	13.4	514	13.0	476	14.4	444	13.8	414	15.2
Other Payer	730	13.8	736	14.6	1,278	13.6	1,318	11.8	1,082	12.3
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	1,947	11.5	1,953	11.5	2,236	11.8	2,474	10.7	2,622	12.0
P02 (NY, NJ, PR)	221	10.5	0	-	0	-	0	-	0	-
P03 (PA, WV, VA, DE, MD, DC)	436	14.6	364	13.9	358	13.8	419	12.5	369	12.5
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	670	9.0	676	9.5	624	11.0	481	11.4	346	11.8
P05 (MN, WI, IL, IN, MI, OH)	1,774	10.6	1,727	11.2	1,251	11.4	1,043	10.3	765	10.6
P06 (NM, OK, AR, LA, TX)	494	12.1	355	14.8	277	17.3	275	17.5	284	16.0
P07 (NE, IA, KS, MO)	265	11.1	228	11.8	250	11.5	169	12.4	212	13.3
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	0	-
P09 (CA, NV, AZ, HI)	0	-	0	-	0	-	0	-	0	-
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

.

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the motor measure is a new measure and has not been used as a stand-alone tool. However, all of the items within the measure are included in a larger instrument (the FIM instrument) which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the motor measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and utility of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Long Term Acute Care Hospitals:

1. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

2. deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

3. Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

> OBJECTIVE: To define the minimal clinically important difference (MCID) for the FIM instrument in patients poststroke. DESIGN: Prospective case series discharged over a 9-month period. SETTING: Long-term acute care hospital. PARTICIPANTS: Patients with stroke (N=113). INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURES: Admission, discharge, and change scores were calculated for the total FIM, motor FIM, and cognitive FIM. Assessments of clinical change were rated at discharge on a 15-point (-7 to +7) Likert scale by attending physicians, with MCID defined at a cutoff score of 3. The FIM change scores associated with MCID were identified from receiver operating characteristic curves. Bayesian analysis was used to determine the probability of individual patients achieving MCID. RESULTS: FIM change scores associated with MCID were 22, 17, and 3 for the total FIM, motor FIM, and cognitive FIM, respectively. The accuracy of the MCID was greater when subjects were categorized based on admission FIM scores than when considering the sample as a whole. Larger FIM change scores were related to MCID in subjects with lower admission FIM scores. CONCLUSIONS: These findings will assist in the interpretation of FIM change scores relative to physicians' assessments of important clinical change.

deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

> OBJECTIVE: To obtain a comprehensive understanding of long-term outcome after severe traumatic brain injury (sTBI). PARTICIPANTS: Forty-six patients with sTBI. DESIGN: Comparison of interdisciplinary evaluation results at discharge from acute care and at 2 to 5 year follow-up. MAIN MEASURES: Extended Glasgow Outcome Scale, the FIM instrument, and the Neurobehavioral Rating Scale-Revised. RESULTS: Significant improvement was observed on the FIM instrument, the Extended Glasgow Outcome Scale, and on 3 factors of the Neurobehavioral Rating Scale-Revised. These measures at discharge were significant predictors of outcome. CONCLUSION: Patients with sTBI 2 to 5 years postinjury showed relatively good physical and functional outcome but poorer cognitive and emotional outcome.

Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

OBJECTIVES: To describe the general characteristics and functional outcomes of individuals treated in a publicly funded, long-term, acquired brain injury rehabilitation program and investigate variables affecting functional outcomes in this patient population. DESIGN: Retrospective database review of demographic, descriptive, and functional outcome assessment data. SETTING: Publicly funded, comprehensive, multidisciplinary, long-term, residential brain injury rehabilitation program in Alberta, Canada (64 beds). PATIENTS: All rehabilitation patients admitted to and discharged from the brain injury program from February 1991 to March 1999 (n = 349). INTERVENTIONS: Multidisciplinary rehabilitation program. MAIN OUTCOME MEASURES: Demographic and descriptive information included sex, age at admission, type and severity of injury, time from injury to long-term program admission, and length of stay (LOS). Functional outcome information included level of care required at admission and discharge, admission and discharge Rappaport disability rating scale scores, and admission and discharge FIM instrument and Functional Assessment Measure scores for a subset of patients. RESULTS: Fifty-nine percent of the subjects had severe traumatic brain injuries (TBI) and 41% had severe nontraumatic brain injuries (NTBI) of various causes. Mean age at admission was older and LOS was longer for NTBI compared with TBI; there were no other differences between the groups in demographic or descriptive measures. The TBI group had significantly lower admission motor subscale scores than the NTBI group, but the groups did not differ on cognitive scores. All functional assessment measures showed statistically significant improvement from admission to discharge, and 85.6% of patients were discharged to community living after a mean LOS of 359.5 days. Functional status at admission, age at admission, length of time between injury and admission, and LOS in the rehabilitation program significantly correlated with functional improvement. CONCLUSIONS: Patients with severe TBI and NTBI who were not candidates for other more conventional forms of rehabilitation showed significant improvement in functional outcomes after extended program admissions. Consideration was also given to the potential insensitivity of commonly used outcome assessment measures in this population.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2776
De.2. **Measure Title:** Functional Change: Change in Motor Score in Long Term Acute Care Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
De.3. **Brief Description of Measure:** Change in rasch derived values of motor function from admission to discharge among adult long term acute care facility patients aged 18 years and older who were discharged alive. The timeframe for the measure is 12 months. The measure includes the following 12 items:Feeding, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
1b.1. **Developer Rationale:** The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or a less intensive setting of care. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing items which are presently collected across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. There are LTACs that are currently collecting data on the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting or less intensive setting upon discharge.

S.4. **Numerator Statement:** Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the LTAC or patients who died within the LTAC are excluded.
S.7. **Denominator Statement:** Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age.
S.10. **Denominator Exclusions:** Patients age at admission less than 18 years old
Patients who died in the LTAC.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Motor_LTAC.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or a less intensive setting of care. Yet the current measures don't adequately capture function or functional improvement. The motor measure is constructed by utilizing items which are presently collected across the post-acute care continuum. Measures of effectiveness, efficiency, timeliness, resource use and safety are an integral part of the items. There are LTACs that are currently collecting data on the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in motor measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in motor function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.

We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting or less intensive setting upon discharge.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

While this is a new measure, UDSMR has historical data on all 12items, and we are able to give information on the measure. See measure evaluation form for the trending data.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

See the measure evaluation sheet for disparity data overtime for the measure.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Development/Wellness, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_Submission-635749865761904393.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

Average change in rasch derived motor functional score from admission to discharge at the facility level for short term rehabilitation patients. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the LTAC or patients who died within the LTAC are excluded.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. The numerator is the average change in rasch derived motor functional score from admission to discharge for each patient at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel,
Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. Impairment type is defined as the primary medical reason for the LTAC stay (such as stroke, joint replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum 12 items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory, Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) at the facility level. Age is the age of the patient at the time of admission to the LTAC. The denominator is meant to reflect the expected motor functional change score at the facility, if the facility had the same distribution of CMGs (based on impairment type, functional status at admission, and age at admission). This adjustment procedure is an indirect standardization procedure (observed facility average/expected facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients age at admission less than 18 years old
Patients who died in the LTAC.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through OASIS.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the CMGs in the excel file provided.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standarization procedure (observed facility average/expected facility average). The numerator is the facility´s average motor functional change score. The denominator is meant to reflect the expected motor functional change score at the facility, if the facility had the same distribution of CMGs (impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all patients during the assessment time frame (12 months).
2. Exclude any patients who died in the LTAC.
3. Exclude any patients who are less than 18 at the time of admission to the LTAC.
3. Calculate the total motor change score for each of the remaining patients (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in excel file).
5. Calculate the average rasch derived motor change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average motor change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average motor change score/facility´s national expected motor change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the appendix.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Long Term Acute Care Hospital
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Motor_Total_LTAC.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Motor Score in Long Term Acute Care Facilities
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation

counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® ("FIM") instrument data from inpatient rehabilitation facilities (IRFs), long term acute care (LTACs), and skilled nursing facilities (SNFs) from the Uniform Data System for Medical Rehabilitation (UDSMR). The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the motor measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2002-2013 were used in examining the data trends over time using the motor measure and patient outcomes of long term acute care facilities.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan

☐ other: Click here to describe	☒ other: patient level/aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term acute care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument (the tool used for this measure) is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 12 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for the long term acute care facility. However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our motor measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

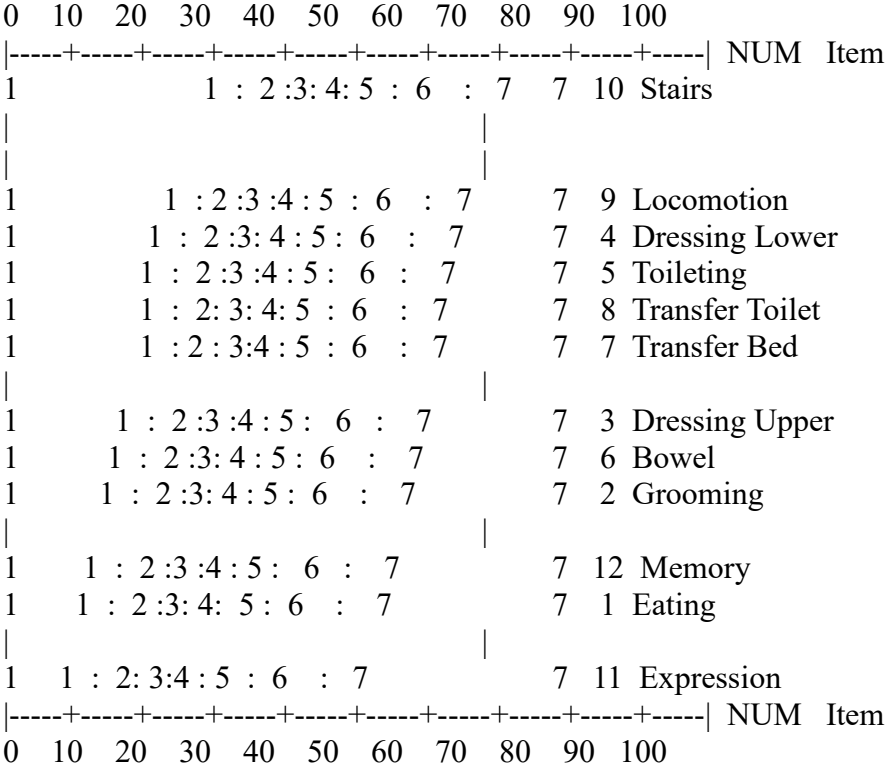
Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (12 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, Memory and Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistant compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.94. The Cronbach Alpha reliability statistic was 0.95. Item correlations within the measure ranged from 0.65 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure, we set the Rasch scale at 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.



The ruler shows that the easiest item is Expression, and the hardest Stairs and that the distances between a level 1 and 2 and 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 12 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
12	.00	17.24	37	37.90	2.28	62	52.00	2.63
13	10.58	8.94	38	38.43	2.27	63	52.73	2.67
14	16.04	6.04	39	38.96	2.26	64	53.47	2.72
15	19.04	4.85	40	39.48	2.25	65	54.25	2.76
16	21.12	4.19	41	40.00	2.24	66	55.05	2.81
17	22.75	3.78	42	40.51	2.23	67	55.88	2.86
18	24.11	3.49	43	41.03	2.23	68	56.74	2.92
19	25.29	3.28	44	41.54	2.23	69	57.64	2.99
20	26.34	3.12	45	42.05	2.23	70	58.58	3.06
21	27.31	2.99	46	42.57	2.23	71	59.57	3.15
22	28.20	2.89	47	43.08	2.24	72	60.63	3.25
23	29.03	2.80	48	43.60	2.25	73	61.76	3.37
24	29.82	2.73	49	44.13	2.26	74	62.98	3.50
25	30.57	2.66	50	44.66	2.28	75	64.30	3.66
26	31.28	2.61	51	45.20	2.29	76	65.75	3.85
27	31.97	2.56	52	45.74	2.31	77	67.37	4.07
28	32.63	2.51	53	46.30	2.34	78	69.19	4.34
29	33.28	2.48	54	46.87	2.36	79	71.29	4.69
30	33.90	2.44	55	47.45	2.39	80	73.79	5.17
31	34.51	2.41	56	48.05	2.42	81	76.91	5.87
32	35.10	2.38	57	48.66	2.45	82	81.15	7.07
33	35.68	2.36	58	49.29	2.49	83	88.16	9.82
34	36.25	2.34	59	49.94	2.52	84	100.00E	17.75
35	36.81	2.32	60	50.61	2.56			
36	37.36	2.30	61	51.29	2.60			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As indicated previously, the reliability of the FIM instrument is well known. The results of the analysis for the measure proposed show the reliability holds even when looking at a subset of FIM instrument items.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the motor measure with the FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new motor measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM® total) explained by the independent variable (motor items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the motor measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM instrument (the gold standard)), and likelihood of discharge to the community setting Linear regression was used to determine functional change, whereas the change in the motor score was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission motor total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the motor measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was highly correlated with the FIM total, 0.932 ($p < 0.001$) and 0.952 ($p < 0.001$), respectively. The correlations remained highly significantly within each venue of care; IRFs, 0.927 ($p < 0.001$) and 0.963 ($p < 0.001$); LTACs, 0.935 ($p < 0.001$) and 0.953 ($p < 0.001$); SNFs, .944 ($p < 0.001$) and .947 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM totals, the r-square values were extremely high, 0.962 and 0.982, respectively. The values remained high at the venue specific level as well; IRFs, 0.945 and 0.974; LTACs, 0.968 and 0.985; SNFs, 0.960 and 0.980.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was very high, 0.866 ($p < 0.001$). In addition, by venue, the correlations remained strong; IRFs, 0.868 ($p < 0.001$); LTACs, 0.887 ($p < 0.001$); SNFs, 0.837 ($p < 0.001$). The linear regression showed high r-squared values as well; all venues, 0.751; IRFs, 0.754; LTACs, 0.786; SNFs, 0.701.

Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in our measure has good predictive ability for discharge setting (community), with a C-statistic of 0.77. By venue, the results are similar; IRFs, 0.75; LTACs, 0.754.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results show high validity across all analyses. The r-squared values were all above 0.8, meaning that the percent of variance explained in the dependent variables by our measure were all more than 80%. In addition, the predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients that died in the post-acute care setting (an unanticipated outcome) and patient under age 18 years, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 1 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives

the expected value based on the case mix and severity mix of each facility. The case mix group classification system groups similarly impaired patients based on functional status at admission or patient severity. This is used for SNFs and IRFs, and the same procedure was applied to the LTAC population. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIMinstrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all CMGs.)

See file uploaded in S.15 for calculations.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified

(*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.
2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.

4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and**

cost of data collection, other feasibility/implementation issues.
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
There are LTACs that are currently using UDSMR and the 12 items in our proposed measure for quality benchmarking, both internally and as a national benchmarking system.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The Functional Change: Change in Motor Score form (this form includes the items for the motor measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) UDSMR www.udsmr.org Quality Improvement (Internal to the specific organization) UDSMR www.udsmr.org

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for LTACs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for skilled nursing facilities, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

The 12 items in our proposed measure are in use in LTACs in the US. Outcomes based on the functional items are currently used for Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and Quality Improvement (Internal to the specific organization).

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
This is a new measure.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences to individuals or populations during the testing of this measure as previously collected data was used.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** Functional_Change_Appendix-635749866379372183.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 03, 2016

Ad.4 **What is your frequency for review/update of this measure?** Unknown, new measure

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

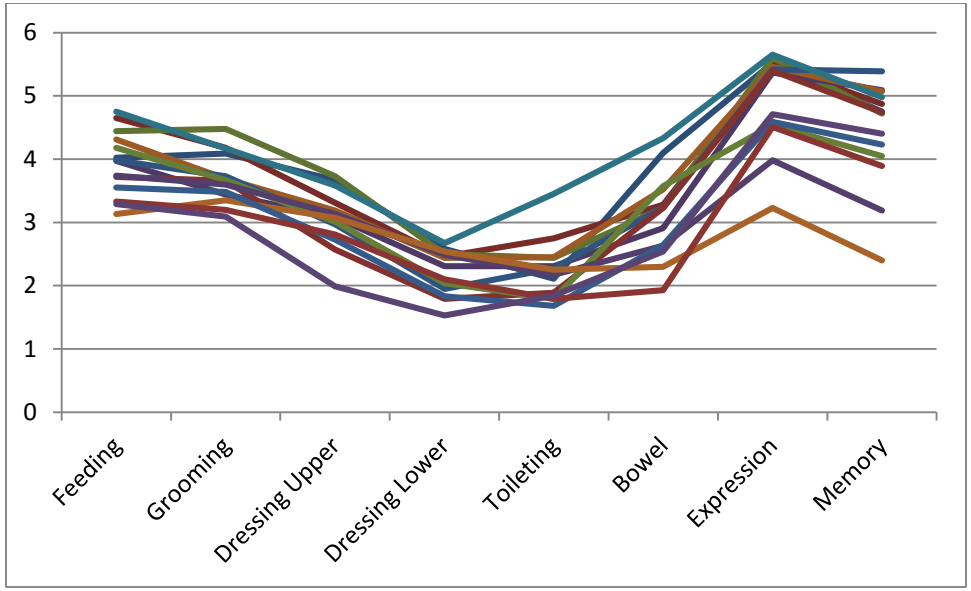
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

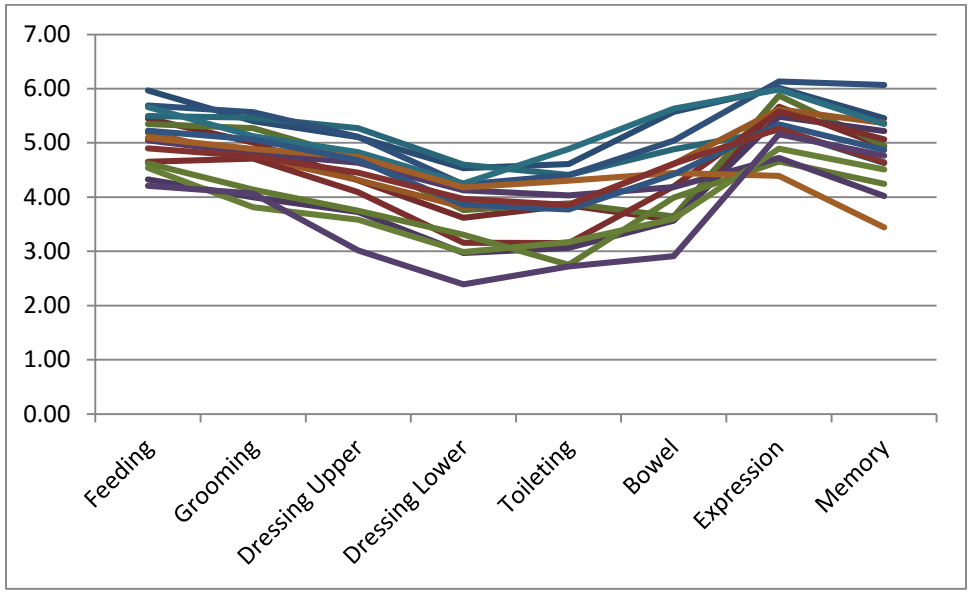
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Long Term Acute Care Facilities, Change in Function: Mobility Measure for Long Term Acute Care Facilities and the Change in Function: Motor Measure for Long Term Acute Care Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

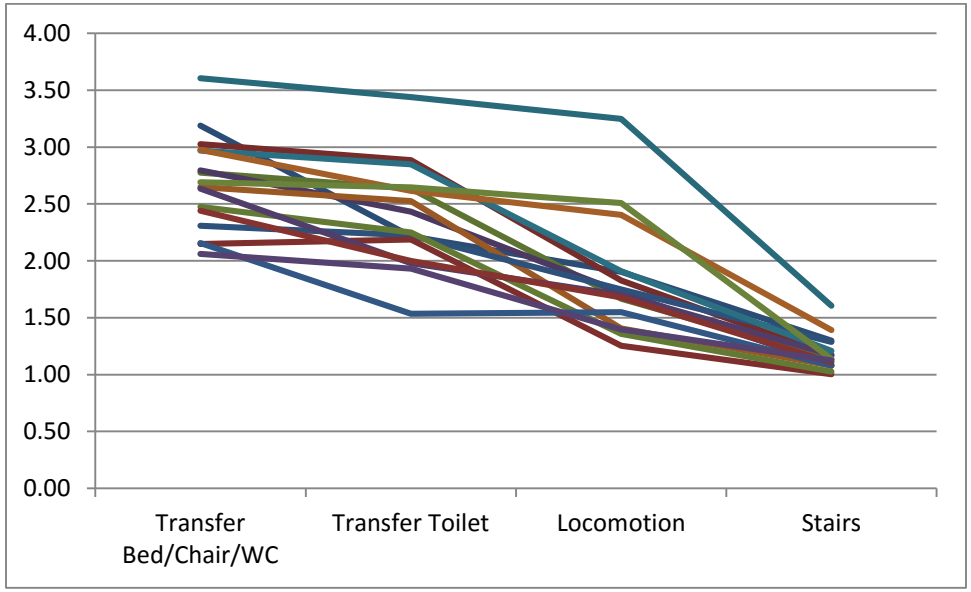
Self-Care Graph: Admission (Year 2009)


6
5
4
3
2
1
0
Feeding
Grooming
Dressing Upper
Dressing Lower
Toileting
Bowel
Expression
Memory

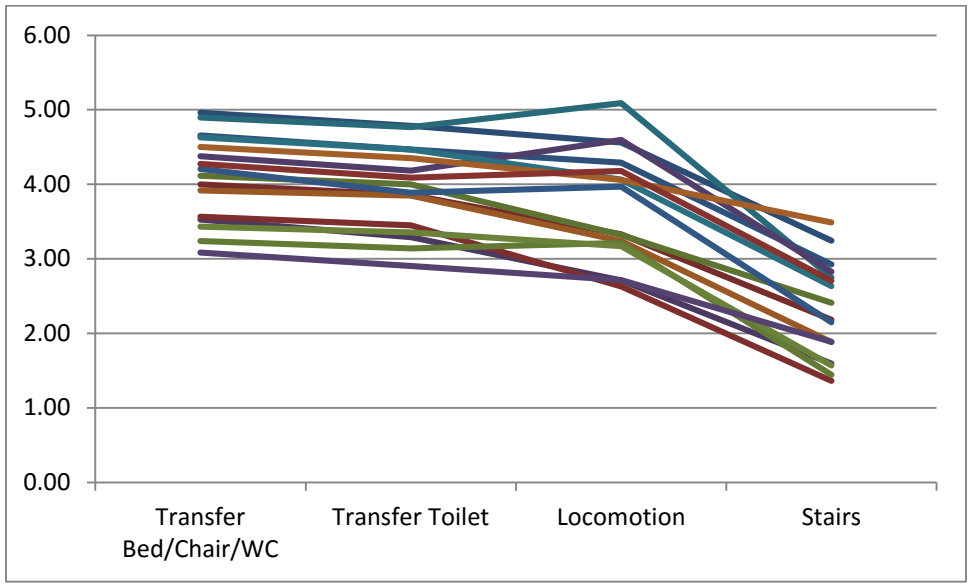
Self-Care Graph Discharge (Year 2009)



Mobility Graph: Admission (Year 2009)



Mobility Graph: Discharge (Year 2009)



Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the LTAC venue of care.

Self-Care Mean Fit Statistics

```
TABLE 3.1 Self Care 8 Items                    ZOU018WS.TXT  Mar 19  9:16 2015
INPUT: 3096 Person  8 Item  REPORTED: 3094 Person  8 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2969 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      36.6       8.0       50.76     3.96       .96    -.1   1.02     .0  |
| S.D.      11.5        .3       13.60     1.46       .71    1.2    .82    1.2  |
| MAX.      55.0       8.0       87.04    10.90      6.32    5.4   8.33    6.2  |
| MIN.       8.0       3.0       11.87     3.00       .05   -3.9    .05   -3.7  |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.60 TRUE SD   12.80  SEPARATION  2.78  Person RELIABILITY  .89  |
|MODEL RMSE   4.22 TRUE SD   12.93  SEPARATION  3.06  Person RELIABILITY  .90  |
| S.E. OF Person MEAN = .25                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     50 Person
  MINIMUM EXTREME SCORE:     75 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      36.2       8.0       50.33     4.59                                |
| S.D.      12.4        .3       16.71     3.40                                |
| MAX.      56.0       8.0      100.06    19.89                                |
| MIN.       8.0       3.0        -.06     3.00       .05   -3.9    .05   -3.7  |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.99 TRUE SD   15.60  SEPARATION  2.61  Person RELIABILITY  .87  |
|MODEL RMSE   5.71 TRUE SD   15.70  SEPARATION  2.75  Person RELIABILITY  .88  |
| S.E. OF Person MEAN = .30                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93
```

Mobility Mean Fit Statistics

```
                                                    03-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      13.8       3.7       31.44    4.51       .94    -.3    .94    -.2 |
| S.D.       6.2        .5       16.49    1.26      1.27    1.4   1.34    1.2 |
| MAX.      27.0       4.0       87.88    9.51      9.90    5.8   9.90    8.5 |
| MIN.       2.0       1.0        8.08    3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89 |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      12.2       3.7       26.70    5.88                                |
| S.D.       6.9        .6       19.75    3.22                                |
| MAX.      28.0       4.0       99.95   13.79                                |
| MIN.       1.0       1.0         .02    3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87 |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

```
TABLE 3.1 All Facilities 12 items                ZOU439WS.TXT  Mar 19  9:43 2015
INPUT: 3096 Person  12 Item  REPORTED: 3094 Person  12 Item  7 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 3013 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      49.2      11.6       45.63     2.83       .99    -.1   1.06     .0 |
| S.D.      17.6        .7       12.31      .98       .67    1.4    .91    1.4 |
| MAX.      83.0      12.0       88.22     9.85      5.13    5.2   9.90    7.7 |
| MIN.      10.0       4.0       10.53     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   3.30 TRUE SD   11.86  SEPARATION  3.59  Person RELIABILITY  .93 |
|MODEL RMSE   2.99 TRUE SD   11.94  SEPARATION  3.99  Person RELIABILITY  .94 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      7 Person
  MINIMUM EXTREME SCORE:     74 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      48.4      11.7       44.66     3.21                               |
| S.D.      18.3        .7       14.26     2.51                               |
| MAX.      84.0      12.0      100.06    17.81                               |
| MIN.      10.0       4.0        -.05     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.30 TRUE SD   13.59  SEPARATION  3.16  Person RELIABILITY  .91 |
|MODEL RMSE   4.07 TRUE SD   13.66  SEPARATION  3.36  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .26                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95
```

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2777
Measure Title: Functional Change: Change in Self Care Score for Long Term Acute Care Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Brief Description of Measure: Change in rasch derived values of self-care function from admission to discharge among adult patients treated in a long term acute care facility who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
Developer Rationale: The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. The self-care measure includes items presently used across the post-acute care continuum. While the items in our proposed measure are not required as part the OASIS system in LTACs, currently there are a number of LTACs that are utilizing the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminating capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care.

Numerator Statement: Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
Denominator Statement: Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age
Denominator Exclusions: Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer states "The primary aim of rehabilitation at LTACs is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or LTAC stay."
- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. While the FIM tool is presently primarily used in inpatient rehabilitation facilities, they state there are LTACs collecting data using the FIM. They provide a list of 3 peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in LTACs.
- The items that comprise the self-care measure are as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory .

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

According to the developer, "The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers."

While this is a new measure, UDSMR has been collecting data on the FIM for more than 20 years so they have historical data to report. The most recent data reported is from 2011 and indicates 23% of cases are below expectation. They offer the following table for LTAC patients:

Year	2007	2008	2009	2010	2011
Selfcare Change Average (Rasch)	14.9	14.7	15.0	14.9	15.0
Case Count	5807	5303	4996	4861	4598
Number of Facilites at or above Expectation (1.0)	9	8	8	9	10
Number of Facilities below Expectation (< 1.0)	9	8	8	5	3
Percent of Facilities at or above Expectation (1.0)	50.0%	50.0%	50.0%	64.3%	76.9%

The developer provided additional documentation stating that the mean score is 36.6, the standard deviation is 11.5, the max is 55.0 and the minimum is 8.0.

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the self care score measure for the years 2007 to 2011. However, information is not provided on whether the differences are statistically significant.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**Yes, the intervention of the patient-goals directed rehabilitation plan within the LTACs is the identified healthcare action.

**Developer provides argument and evidence supporting measurement of basic ADL functional status outcomes for LTACs. Agree this measure passes this criterion.

**The developers' evidence is from the existing FIM instrument. They are using a sum of 8 scales: : Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.

1b. Performance Gap

Comments:

**Moderate rating of performance gap in this measure. The historical data shows a trending improvement beginning in 2010, leaving room for doubt as to the number of facilities still operating below expectations in 2016.

**Deferred as performance gap unclear to me based upon my review. I appreciate that the developer has provided weak/negative ICCs noted in their across facility reliability analyses but I cannot determine if this is signifies a clinically meaningful performance gap without seeing actual distribution of facility-level results.

**Compared to the mobility measure, the variation regionally and by payer source is not as pronounced. However, if this measure is rolled out to a larger group (rather than the self-selected group which chose to participate in using this measure), there might be higher variability expected.

1c. PRO-PM

Comments:

**N/A (which is one possible limitation of this measure, that it is not directly asking patients about function, but this may not be feasible for this specific patient population)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, OASIS

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:
  - Average change in rasch derived self-care functional score from admission to discharge at the facility level for short term rehabilitation patients, over Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age.
  - Average is calculated as (sum of change at the patient level/total number of patients).

- The calculation algorithm is included.
- Patients under age 18 and patients who died in the LTAC are excluded.
- A data dictionary is included.
- The measure is stratified by risk category using an indirect standardization procedure (observed facility average/expected facility average)

Questions for the Committee :

- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented using inter and intra-rater reliability
  - This measure uses a subset of the FIM, so a Rasch analysis was conducted to test the following:
    - the psychometric properties of the subset of 8 items within the three venues of post-acute care, IRFs, LTACs, and SNFs
    - the measure reliability at both the person and item level
    - to determine the fit of each item within the measure (8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) through infit and outfit statistics and item specific correlations.
- Internal consistency is demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but is working to do so prior to the in-person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities (perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities. UDSMR has indicated they are working to fulfill these data needs.*

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.70 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm

Precise specifications – yes (box 1) -> empirical testing of data elements (box 2) -> TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- The developers examined the concurrent validity of the self-care measure with the total FIM total score, both at admission and discharge. The two tests of validity used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (self-care items). The admission and discharge values were examined separately.
- Predictive validity of the self-care score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:

- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for LTACs are 0.928 ($p < 0.001$) at admission and 0.888 ($p < 0.001$) at discharge. For predictive validity, LTACs scored 0.757 ($p < 0.001$).
- The developer indicated that r-square values were very high for admission FIM total and discharge FIMR total. For LTACs: .861 and 0.788;
- The linear regression showed significant, high r-squared values as well; all venues, 0.519 and for LTACs, 0.574

- For all venues, the logistic regression analysis shows that the gain in self-care has good predictive ability for discharge setting (community), with a C-statistic of 0.76. By venue, the results are LTACs, 0.73.
- The developer summarized the results as follows: *The results show the self-care measure is valid; the measure demonstrated construct, concurrent, discriminant and predictive validity in all analyses. The r-square values were all consistent, 0.6 or higher, meaning that the percent of variance explained in the dependent variables by our measure were all more than 60%. The predictive validity was also high.*

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the facility were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

No information provided

Question for the Committee:

Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**I support the validity of the specifications with the evidence.

**Additional information provided by developer demonstrates an ICC of 0.64 using a split sample test-retest method within facilities. It would be helpful to know from the developer if their 30 patient per facility sample for this testing was the split or total sample (ie, 30 patients per sample or only 15) and what the average sample size for facilities is to help interpret this data further. The testing results are robust but I would like to know the facility-level patient volume before grading this criterion.

**Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age. Probably will require the individual facilities to join a 3rd party to provide these calculations. Federal Register (found on USDSMR website): Greater variation in discharge to community rates is seen in the SNF setting, with rates ranging from 31 to 65 percent. A multi-center study of 23 LTCHs demonstrated that 28.8 percent of 1,061 patients who were ventilator-dependent on admission were discharged to home. A single-center study revealed that 31 percent of LTCH hemodialysis patients were discharged to home. One study noted that 64 percent of beneficiaries who were discharged from the home health episode did not use any other acute or post-acute services paid by Medicare in the 30 days after discharge.48 However, significant numbers of patients were admitted to hospitals (29 percent) and lesser numbers to SNFs (7.6 percent), IRFs (1.5 percent), home health (7.2 percent) or hospice (3.3 percent).

2a2. Reliability Testing

Comments:

**A facility-level computed performance measure was assessed for the data elements. I cannot professionally speak to the appropriateness of the methods used. But seems that both the sample size and results support strong reliability of the measure. I believe variability will occur between individuals more so than at the facility level.

**Indirect standardization procedure. Internal consistency was demonstrated.

2b.2 Validity Testing

Comments:

**It seems to me that the validity of the measure is closely tied to the validity of the FIM tool. But the validity of the measure was combined between venues. Shouldn't they look at validity of the measure between venues?

**The developer reports concurrent and predictive validity testing. The predictive testing assessed prediction of functional gain assessed by Total FIM as well as discharge to the community setting. The concurrent validity tested the correlation between the self-care measure (derived from a subset of the total FIM) and the total FIM. I cannot give credit for empiric validity testing using the original long form as a gold standard when the measure is derived from this

instrument. This analysis shows that the data element of the self care survey is valid in comparison to the FIM, but does not demonstrate validity of the measure result. However, the predictive validity of the association of the measure result with discharge to the community does provide assurance of measure validity in my opinion.

Notably missing is any commentary or input from patients on the validity of the self care instrument items, although this patient population may be challenging to study due to their clinical limitations.

**The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for LTACs are 0.928 ($p < 0.001$) at admission and 0.888 ($p < 0.001$) at discharge. For predictive validity, LTACs scored 0.757 ($p < 0.001$).

The developer indicated that r-square values were very high for admission FIM total and discharge FIMR total. For LTACs: .861 and 0.788;

The linear regression showed significant, high r-squared values as well; all venues, 0.519 and for LTACs, 0.574

For all venues, the logistic regression analysis shows that the gain in self-care has good predictive ability for discharge setting (community), with a C-statistic of 0.76. By venue, the results are LTACs, 0.73.

2b3.-2b7. Test Related to Potential Threats

Comments:

**I support the logic used for exclusions and risk adjustments. Since all data elements are required, missing data should not be a threat to validity. These exclusions and risk adjustments seem to be a standard process. And risk adjustments in this measure may very well be necessary.

**The developer does not provide data (or I could not find it) regarding the number of missing surveys or capture rate across facilities. Although the exclusions are minimal (patients under 18 and those patient that die in the facility), the issue of capture rate and representativeness is not clear.

If the developer can address some of the gaps in knowledge about capture rate and representativeness of the measure sample of the facility's patients as a whole, I agree with the NQF staff's moderate validity rating.

**Exclusions are persons under 18 and death.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- The developer reports there are LTACs currently using the FIM
- Commercial use requires a license agreement and has a fee. The developer reports the following: "The Functional Change: Change in Motor Score form (this form includes the items for the self-care measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

○ *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**All the data elements are collected during care delivery (or should be) - but the copyright on external use of the FIM tool itself (and representative score) may inhibit some feasibility to its implementation. The self-care measure's feasibility is closely tied to the availability and respective fees of the FIM tool.

**The FIM is not universally collected and not part of the mandatory OASIS data collection for LTACs. While the developer indicates that many facilities are currently collecting this, supporting its global feasibility, there is clearly an added burden to the facility to collect additional data beyond what they are already collecting.

This is also a physician/provider-reported measure, requiring additional resources to collect the data required for the measure.

Further, the developer notes that the FIM and self care instrument are copyrighted but there is not cost to use them for internal auditing. It is unclear if this exemption includes federal reporting or payment programs.

I defer making a recommendation about feasibility until discussion with the committee and NQF staff.

**Training of personnel in the LTACH's to perform FIM rating will be necessary, which does add an expense as the UDSMR requires all individuals who are giving ratings to undergo training. This hasn't been clarified by the developer as to potential burden.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

The developer reports that "Currently UDSMR provides both internal reporting and national benchmarking for LTACs who subscribe to the UDSMR software/outcomes reporting."

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results

- New measure – not available

Unexpected findings (positive or negative) during implementation

- None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**The measure is not publicly reported. But it's results could be used to indicate a facilities quality of rehabilitation of self-care. Patients seeking to reduce the amount of assistance needed upon discharge from LTACs could benefit from this measurement.

**This measure is currently in use by the UDSMR, supporting its usability. Provision of a description of how the information is shared with facilities would assist in evaluating its usability. Is the data presented in relation to national data or in isolation? Some contextual information about how this information can be interpreted by facilities would be beneficial.

Given that it is in reporting, it is likely to pass moderate usability in my opinion, although additional detail would be helpful.

**No details on public reporting of the measure. Again, they require users to have a subscription/license through UDSMR. Any public reporting isn't described on their website.

Criterion 5: Related and Competing Measures

Related or competing measures

None reported

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Self Care Score for Long Term Acute Care Facilities
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus
The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:
Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

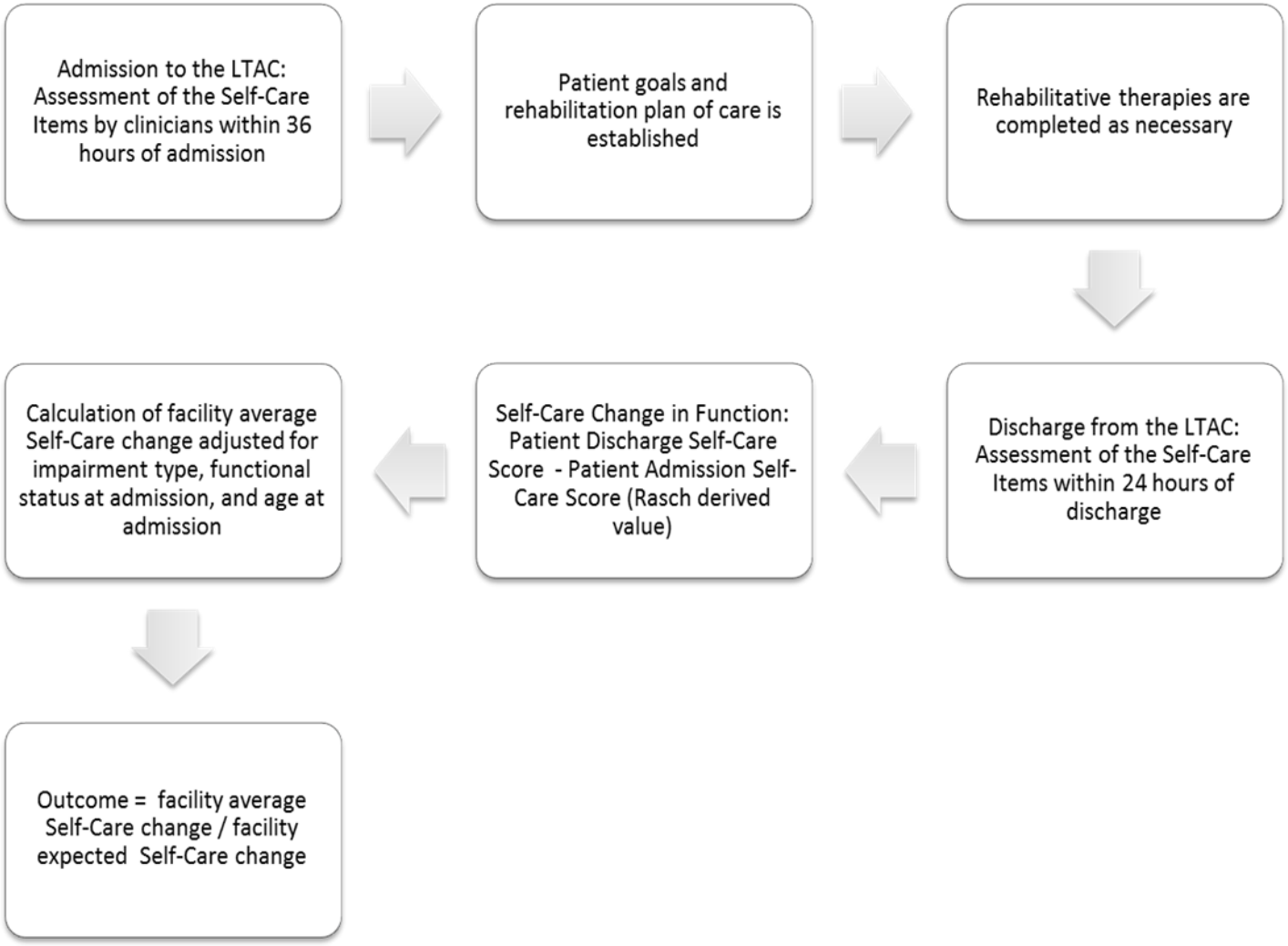
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Long Term Acute Care Hospitals (LTACs) are one part of a multi-level post-acute care continuum. The primary aim of rehabilitation at LTACs is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or LTAC stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are LTACs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for patients seen at LTACs. To date, the self-care measure has not been reported on as a stand-alone measure. However, the items of the self-care measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument. The self-care measure is intended to be administered within 24 hours of the patient's admission to the IRF and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the self-care measure are as follows: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory). All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for LTAC patients:



UDSMR has been a data repository for the FIM instrument among LTAC patients, of which the items of the self-care measure are nested within for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the self-care measure for the years 2007 to 2011 for LTAC patients:

Year	2007	2008	2009	2010	2011
Selfcare Change Average (Rasch)	14.9	14.7	15.0	14.9	15.0
Case Count	5807	5303	4996	4861	4598
Number of Facilites at or above Expectation (1.0)	9	8	8	9	10
Number of Facilities below Expectation (< 1.0)	9	8	8	5	3
Percent of Facilities at or above Expectation (1.0)	50.0%	50.0%	50.0%	64.3%	76.9%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the self-care measure for the years 2007 to 2011:

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2007		2008		2009		2010		2011	
	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)	Case Count	Selfcare Change Average (Rasch)
Gender										
Male	3,126	14.8	2,897	14.8	2,724	14.8	2,641	14.9	2,493	15.0
Female	2,676	15.1	2,398	14.6	2,267	15.2	2,215	14.7	2,101	14.9
Ethnicity										
White	4,653	15.2	4,346	15.0	3,895	15.1	3,606	14.9	3,508	14.9
Black	636	14.3	547	13.3	538	13.9	463	14.2	379	14.5
Hispanic	62	13.2	61	14.6	56	14.9	81	15.7	47	14.5
Other Ethnicity	456	13.4	349	13.6	507	15.5	711	15.1	664	15.6
Payer Source										
Medicare	3,444	14.6	3,075	14.3	2,264	14.4	2,222	14.1	2,342	14.4
Medicaid	366	14.7	337	14.2	321	14.6	246	14.3	225	14.5
Commercial	679	14.8	641	14.9	657	14.9	631	15.0	535	14.8
Blue Cross	588	15.9	514	16.0	476	15.6	444	15.8	414	16.4
Other Payer	730	15.7	736	15.7	1,278	15.9	1,318	15.7	1,082	15.8
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	1,947	16.3	1,953	16.1	2,236	15.9	2,474	15.5	2,622	15.4
P02 (NY, NJ, PR)	221	17.2	0	-	0	-	0	-	0	-
P03 (PA, WV, VA, DE, MD, DC)	436	14.2	364	13.6	358	13.8	419	13.7	369	13.6
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	670	14.2	676	14.3	624	15.7	481	16.4	346	16.7
P05 (MN, WI, IL, IN, MI, OH)	1,774	13.5	1,727	13.5	1,251	13.5	1,043	13.1	765	13.6
P06 (NM, OK, AR, LA, TX)	494	14.6	355	14.3	277	14.1	275	14.3	284	14.0
P07 (NE, IA, KS, MO)	265	15.6	228	15.6	250	15.3	169	15.4	212	15.3
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	0	-
P09 (CA, NV, AZ, HI)	0	-	0	-	0	-	0	-	0	-
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the self-care measure is a new measure and has not been used as a stand-alone tool. However, all of the items within the measure are included in a larger instrument (the FIM instrument) which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the self-care measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and utility of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Long Term Acute Care Hospitals:

1. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

2. deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

3. Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE
1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE
1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE
1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE
1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE
1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE
If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

OBJECTIVE: To define the minimal clinically important difference (MCID) for the FIM instrument in patients poststroke. DESIGN: Prospective case series discharged over a 9-month period. SETTING: Long-term acute care hospital. PARTICIPANTS: Patients with stroke (N=113). INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURES: Admission, discharge, and change scores were calculated for the total FIM, motor FIM, and cognitive FIM. Assessments of clinical change were rated at discharge on a 15-point (-7 to +7) Likert scale by attending physicians, with MCID defined at a cutoff score of 3. The FIM change scores associated with MCID were identified from receiver operating characteristic curves. Bayesian analysis was used to determine the probability of individual patients achieving MCID. RESULTS: FIM change scores associated with MCID were 22, 17, and 3 for the total FIM, motor FIM, and cognitive FIM, respectively. The accuracy of the MCID was greater when subjects were categorized based on admission FIM scores than when considering the sample as a whole. Larger FIM change scores were related to MCID in subjects with lower admission FIM scores. CONCLUSIONS: These findings will assist in the interpretation of FIM change scores relative to physicians' assessments of important clinical change.

deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

OBJECTIVE: To obtain a comprehensive understanding of long-term outcome after severe traumatic brain injury (sTBI). PARTICIPANTS: Forty-six patients with sTBI. DESIGN: Comparison of interdisciplinary evaluation results at discharge from acute care and at 2 to 5 year follow-up. MAIN MEASURES: Extended Glasgow Outcome Scale, the FIM instrument, and the Neurobehavioral Rating Scale-Revised. RESULTS: Significant improvement was observed on the FIM instrument, the Extended Glasgow Outcome Scale, and on 3 factors of the Neurobehavioral Rating Scale-Revised. These measures at discharge were significant predictors of outcome. CONCLUSION: Patients with sTBI 2 to 5 years postinjury showed relatively good physical and functional outcome but poorer cognitive and emotional outcome.

Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

OBJECTIVES: To describe the general characteristics and functional outcomes of individuals treated in a publicly funded, long-term, acquired brain injury rehabilitation program and investigate variables affecting functional outcomes in this patient population. DESIGN: Retrospective database review of demographic, descriptive, and functional outcome assessment data. SETTING: Publicly funded, comprehensive, multidisciplinary, long-term, residential brain injury rehabilitation program in Alberta, Canada (64 beds). PATIENTS: All rehabilitation patients admitted to and discharged from the brain injury program from February 1991 to March 1999 (n = 349). INTERVENTIONS: Multidisciplinary rehabilitation program. MAIN OUTCOME MEASURES: Demographic and descriptive information included sex, age at admission, type and severity of injury, time from injury to long-term program admission, and length of stay (LOS). Functional outcome information included level of care required at admission and discharge, admission and discharge Rappaport disability rating scale scores, and admission and discharge FIM instrument and Functional Assessment Measure scores for a subset of patients. RESULTS: Fifty-nine percent of the subjects had severe traumatic brain injuries (TBI) and 41% had severe nontraumatic brain injuries (NTBI) of various causes. Mean age at admission was older and LOS was longer for NTBI compared with TBI; there were no other differences between the groups in demographic or descriptive measures. The TBI group had significantly lower admission motor subscale scores than the NTBI group, but the groups did not differ on cognitive scores. All functional assessment measures showed statistically significant improvement from admission to discharge, and 85.6% of patients were discharged to community living after a mean LOS of 359.5 days. Functional

status at admission, age at admission, length of time between injury and admission, and LOS in the rehabilitation program significantly correlated with functional improvement. CONCLUSIONS: Patients with severe TBI and NTBI who were not candidates for other more conventional forms of rehabilitation showed significant improvement in functional outcomes after extended program admissions. Consideration was also given to the potential insensitivity of commonly used outcome assessment measures in this population.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2777
De.2. **Measure Title:** Functional Change: Change in Self Care Score for Long Term Acute Care Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
De.3. **Brief Description of Measure:** Change in rasch derived values of self-care function from admission to discharge among adult patients treated in a long term acute care facility who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
1b.1. **Developer Rationale:** The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. The self-care measure includes items presently used across the post-acute care continuum. While the items in our proposed measure are not required as part the OASIS system in LTACs, currently there are a number of LTACs that are utilizing the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminating capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care.

S.4. **Numerator Statement:** Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.
S.7. **Denominator Statement:** Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age
S.10. **Denominator Exclusions:** Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Self_Care_LTAC-635950315917865122.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. The self-care measure includes items presently used across the post-acute care continuum. While the items in our proposed measure are not required as part the OASIS system in LTACs, currently there are a number of LTACs that are utilizing the items for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in self-care measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminating capabilities for rehabilitation patients, and predictive of change in self-care function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see Measure Evaluation Form for data over time

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see Measure Evaluation Form for disparities data

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;

OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Development/Wellness, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_Submission_Self_Care-635749886179500305.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Average change in rasch derived self-care functional score from admission to discharge at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 Months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. The numerator is the average change in rasch derived self-care functional score from admission to discharge for each patient at the facility level, including items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory. Average is calculated as: (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted expected change in rasch derived values, adjusted for CMG (Case Mix Group), based on impairment type, admission functional status, and age

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. Impairment type is defined as the primary medical reason for the LTAC stay (such as stroke, joint replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum 8 self-care items ((Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) at the facility level. Age is the age of the patient at the time of admission to the LTAC. The denominator is meant to reflect the expected Self-Care functional change score at the facility, if the facility had the same distribution of CMGs (based on impairment type, functional status at admission, and age at admission). This adjustment procedure is an indirect standardization procedure (observed facility average/expected facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through OASIS.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the CMGs in the excel file provided.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standarization procedure (observed facility average/expected facility average). The numerator is the facility´s average self-care functional change score. The denominator is meant to reflect the expected Self-Care functional change score at the facility, if the facility had the same distribution of CMGs(impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all patients during the assessment time frame (12 months).
2. Exclude any patients who died in the LTAC.
3. Exclude any patients who are less than 18 at the time of admission to the LTAC.
3. Calculate the total self-care change score for each of the remaining patients (sum of change at the patient level for all items (Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in excel file).
5. Calculate the average rasch derived self-care change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average self-care change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average self-care change score/facility´s national expected self-care change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the appendix.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Long Term Acute Care Hospital
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Self_Care_LTAC.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Self Care Score for Long Term Acute Care Facilities Click here to enter measure title

Date of Submission: 3/31/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the

information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® ("FIM") instrument data from inpatient rehabilitation facilities (IRFs), long term acute care (LTACs), and skilled nursing facilities (SNFs) from the Uniform Data System for Medical Rehabilitation (UDSMR). The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the self-care measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2010 - 2014 were used in examining the data trends over time using the self-care measure and patient outcomes of long term acute care facilities.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan

☐ other: Click here to describe	☒ other: patient level, aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term acute care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument (the tool used for this measure), is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 8 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for the long term acute care facilities.However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our self-care measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

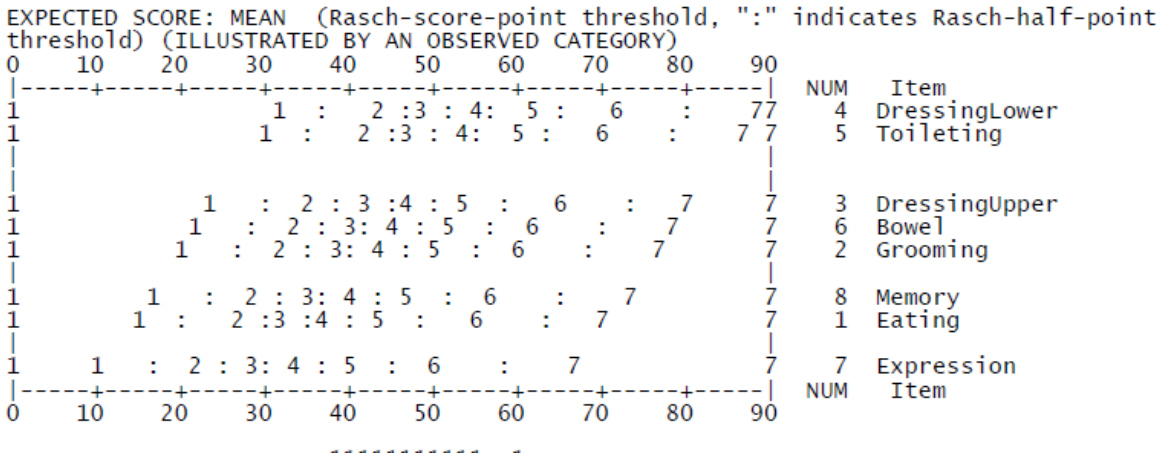
Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (8 items: Eating, Grooming, Dressing Upper Body, Dressing Lower Body, Toileting, Bowel, Expression, and Memory) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistant compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.70 to 0.84. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure the Rasch scale was set to 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.



The ruler shows that the easiest functional item is Expression, and the most challenging functional item is Dressing Lower, additionally, the distances between a level 1 and 2 and 5, 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 8 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
8	.00E	19.37	25	38.91	3.03	42	54.11	3.50
9	11.92	10.11	26	39.75	3.01	43	55.26	3.61
10	18.17	6.88	27	40.58	3.00	44	56.50	3.72
11	21.64	5.55	28	41.40	3.00	45	57.81	3.86
12	24.08	4.81	29	42.22	3.00	46	59.23	4.01
13	25.99	4.35	30	43.05	3.00	47	60.77	4.18
14	27.60	4.03	31	43.88	3.01	48	62.44	4.37
15	29.00	3.80	32	44.71	3.02	49	64.28	4.60
16	30.26	3.62	33	45.56	3.04	50	66.33	4.88
17	31.42	3.49	34	46.41	3.06	51	68.67	5.24
18	32.50	3.38	35	47.28	3.09	52	71.42	5.73
19	33.52	3.29	36	48.17	3.12	53	74.81	6.48
20	34.49	3.22	37	49.07	3.16	54	79.39	7.78
21	35.42	3.16	38	50.00	3.21	55	86.98	10.87
22	36.33	3.12	39	50.97	3.27	56	100.00E	19.83
23	37.20	3.08	40	51.97	3.34			
24	38.07	3.05	41	53.01	3.41			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*).
As indicated previously, the reliability of the FIM instrument is well known. The results of the analysis for the measure proposed show the reliability holds even when looking at a subset of FIM instrument items.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Performance measure score**
☒ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or

resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the self-care measure with the total FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new self-care measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (self-care items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the self-care measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM® instrument (the gold standard)), and likelihood of discharge to the community setting Linear regression was used to determine functional change, whereas the change in self-care was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission self-care total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the self-care measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was highly correlated with the FIM total, 0.929 ($p < 0.001$) and 0.881 ($p < 0.001$), respectively. The correlations remained significant within each venue of care; IRFs, 0.933 ($p < 0.001$) and 0.896 ($p < 0.001$); LTACs, 0.928 ($p < 0.001$) and 0.888 ($p < 0.001$); SNFs, 0.937 ($p < 0.001$) and 0.871 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM totals, the r-square values were very high for admission FIM total and discharge FIM total, 0.864 and 0.775, respectively. The values remained similar at the venue specific level as well; IRFs, 0.870 and 0.804; LTACs, 0.861 and 0.788; SNFs, 0.877 and 0.758.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was strong, 0.721 ($p < 0.001$). In addition, by venue, the correlations remained strong; IRFs, 0.780 ($p < 0.001$); LTACs, 0.757 ($p < 0.001$); SNFs, 0.681 ($p < 0.001$). The linear regression showed significant, high r-squared values as well; all venues, 0.519; IRFs, 0.608; LTACs, 0.574; SNFs, 0.464.

Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in self-care has good predictive ability for discharge setting (community), with a C-statistic of 0.76. By venue, the results are similar; IRFs, 0.74; LTACs, 0.73; SNFs, 0.80.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results show the self-care measure is valid; the measure demonstrated construct, concurrent, discriminant and predictive validity in all analyses. The r-square values were all consistent, 0.6 or higher, meaning that the percent of variance explained in the dependent variables by our measure were all more than 60%. The predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients that died in the post-acute care setting (an unanticipated outcome) and patient aged 18 years and older, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

No statistical tests completed.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 1 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The case mix group classification system groups similarly impaired patients based on functional status at admission or patient severity. This is used for SNFs and IRFs, and the same procedure will be applied to the LTACs. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIM instrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all CMGs.)

See file uploaded in S.15 for calculations.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.
2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.
4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
A number of LTACs currently use UDSMR and collect data on the items in our proposed measure for quality benchmarking, both internally and as a national benchmarking system. Therefore the measure is feasible for use, demonstrates low administrative burden and has no implementation issues.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The Functional Change: Change in Motor Score form (this form includes the items for the self-care measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) UDSMR www.udsmr.org Quality Improvement (Internal to the specific organization) UDSMR www.udsmr.org

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for LTACs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for LTACs, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

Outcomes included the items in our proposed measure are currently used for Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and Quality Improvement (Internal to the specific organization).

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences to individuals or populations during the testing of this measure as previously collected data was used.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Functional_Change_Appendix-635749891008549447.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr,org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 03, 2016

Ad.4 **What is your frequency for review/update of this measure?** Unknown, new measure

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

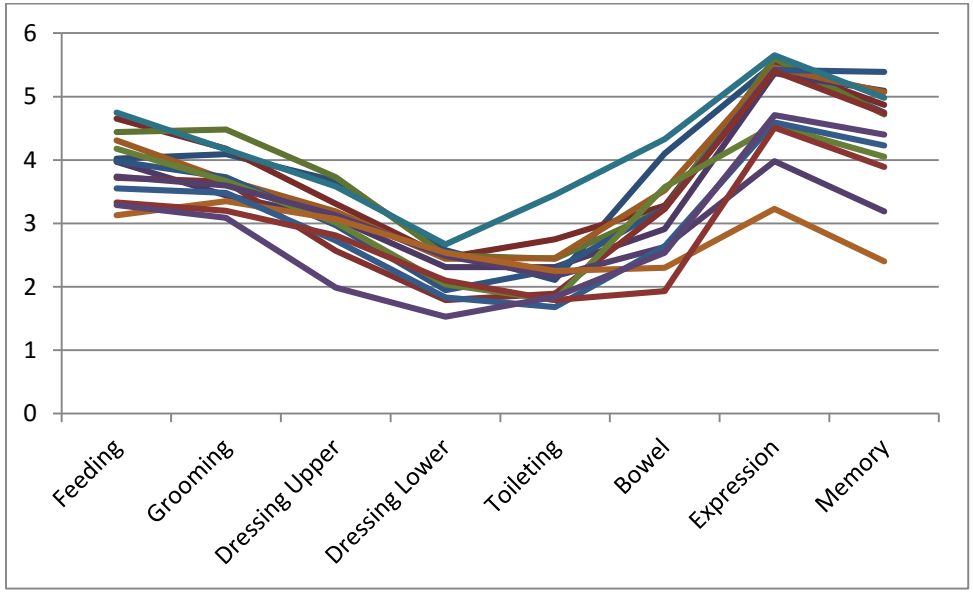
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

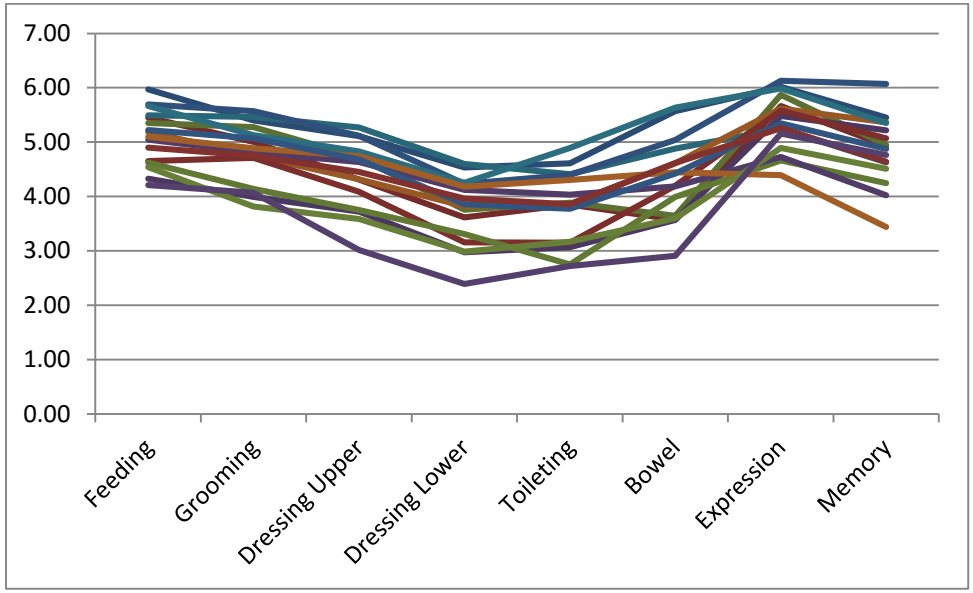
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Long Term Acute Care Facilities, Change in Function: Mobility Measure for Long Term Acute Care Facilities and the Change in Function: Motor Measure for Long Term Acute Care Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

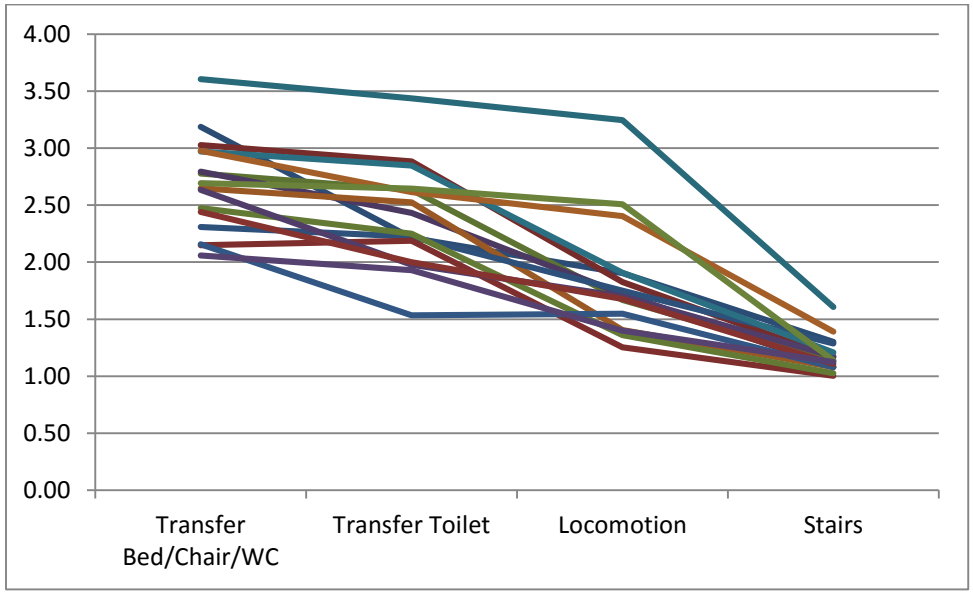
Self-Care Graph: Admission (Year 2009)



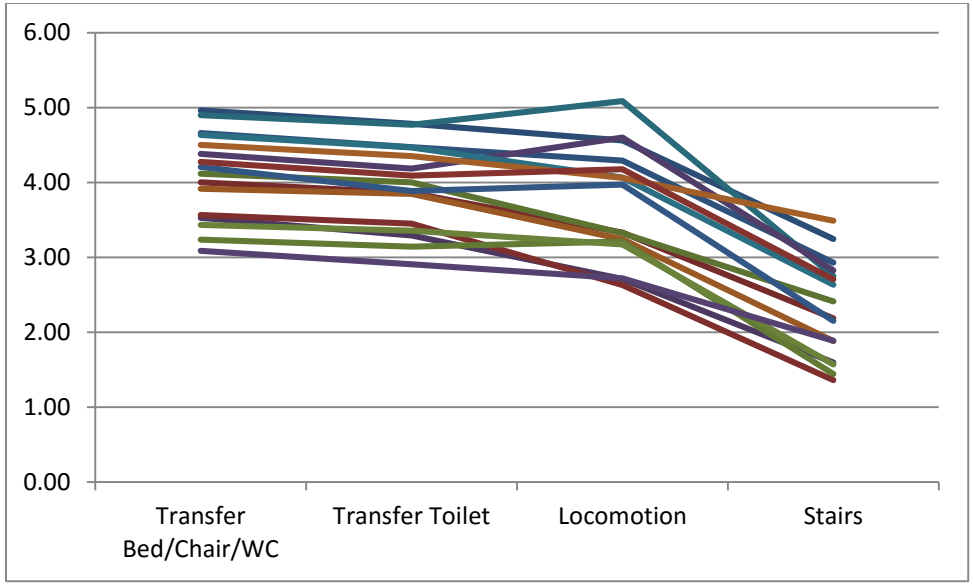
Self-Care Graph Discharge (Year 2009)



Mobility Graph: Admission (Year 2009)



Mobility Graph: Discharge (Year 2009)


6.00
5.00
4.00
3.00
2.00
1.00
0.00
Transfer Bed/Chair/WC
Transfer Toilet
Locomotion
Stairs

Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the LTAC venue of care.

Self-Care Mean Fit Statistics

TABLE 3.1 Self Care 8 Items ZOU018WS.TXT Mar 19 9:16 2015
INPUT: 3096 Person 8 Item REPORTED: 3094 Person 8 Item 7 CATS WINSTEPS 3.73

SUMMARY OF 2969 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.6	8.0	50.76	3.96	.96	-.1	1.02	.0
S.D.	11.5	.3	13.60	1.46	.71	1.2	.82	1.2
MAX.	55.0	8.0	87.04	10.90	6.32	5.4	8.33	6.2
MIN.	8.0	3.0	11.87	3.00	.05	-3.9	.05	-3.7

REAL RMSE	4.60	TRUE SD	12.80	SEPARATION 2.78	Person RELIABILITY .89
MODEL RMSE	4.22	TRUE SD	12.93	SEPARATION 3.06	Person RELIABILITY .90
S.E. OF Person MEAN = .25					

MAXIMUM EXTREME SCORE: 50 Person
MINIMUM EXTREME SCORE: 75 Person
LACKING RESPONSES: 2 Person

SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.2	8.0	50.33	4.59				
S.D.	12.4	.3	16.71	3.40				
MAX.	56.0	8.0	100.06	19.89				
MIN.	8.0	3.0	-.06	3.00	.05	-3.9	.05	-3.7

REAL RMSE	5.99	TRUE SD	15.60	SEPARATION 2.61	Person RELIABILITY .87
MODEL RMSE	5.71	TRUE SD	15.70	SEPARATION 2.75	Person RELIABILITY .88
S.E. OF Person MEAN = .30					

Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93

Mobility Mean Fit Statistics

```
                                                  03-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       13.8       3.7       31.44     4.51       .94    -.3    .94    -.2 |
| S.D.        6.2        .5       16.49     1.26      1.27    1.4   1.34    1.2 |
| MAX.       27.0       4.0       87.88     9.51      9.90    5.8   9.90    8.5 |
| MIN.        2.0       1.0        8.08     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89 |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                          MODEL         INFIT        OUTFIT    |
|           SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       12.2       3.7       26.70     5.88                              |
| S.D.        6.9        .6       19.75     3.22                              |
| MAX.       28.0       4.0       99.95    13.79                              |
| MIN.        1.0       1.0         .02     3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87 |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

```
TABLE 3.1 All Facilities 12 items              ZOU439WS.TXT  Mar 19  9:43 2015
INPUT: 3096 Person  12 Item  REPORTED: 3094 Person  12 Item  7 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 3013 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      49.2      11.6       45.63    2.83       .99    -.1   1.06     .0 |
| S.D.      17.6        .7       12.31     .98       .67    1.4    .91    1.4 |
| MAX.      83.0      12.0       88.22    9.85      5.13    5.2   9.90    7.7 |
| MIN.      10.0       4.0       10.53    2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   3.30 TRUE SD   11.86  SEPARATION  3.59  Person RELIABILITY  .93 |
|MODEL RMSE   2.99 TRUE SD   11.94  SEPARATION  3.99  Person RELIABILITY  .94 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      7 Person
  MINIMUM EXTREME SCORE:     74 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      48.4      11.7       44.66    3.21                                |
| S.D.      18.3        .7       14.26    2.51                                |
| MAX.      84.0      12.0      100.06   17.81                                |
| MIN.      10.0       4.0        -.05    2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.30 TRUE SD   13.59  SEPARATION  3.16  Person RELIABILITY  .91 |
|MODEL RMSE   4.07 TRUE SD   13.66  SEPARATION  3.36  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .26                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95
```

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2778
Measure Title: Functional Change: Change in Mobility Score for Long Term Acute Care Facilities
Measure Steward: Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc. and its successor in interest, UDSMR, LLC.
Brief Description of Measure: Change in rasch derived values of mobility function from admission to discharge among adult LTAC patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 4 mobility items:Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
Developer Rationale: The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. There are LTACs that are currently collect data on the items in the proposed measure for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community. The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care after their LTAC stay.

Numerator Statement: Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.
Denominator Statement: Facility adjusted adjusted expected change in rasch derived values, adjusted at the Case Mix Group level.
Denominator Exclusions: Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developers provide a flow chart linking the completion of rehabilitation therapy to the outcome of facility improvement in scores. They provide a list of three peer-reviewed journal articles that demonstrate validity and use of the FIM instrument in LTACs.
- In addition, they provide summaries/abstracts from three articles that support the following: *The primary aim of rehabilitation is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or SNF stay.*
- The items in the mobility measure are Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- According to the developer, "The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers."
- This is a new measure, but UDSMR has been collecting data on the FIM instrument for 20 years, so they are able to report on trends. The most recent data available is from 2011, where more than half (54%) of facilities were below expectation:

Year	2007	2008	2009	2010	2011
Mobility Change Average (Rasch)	18.1	18.8	19.0	18.2	19.8
Case Count	5807	5303	4996	4861	4598
Number of Facilites at or above Expectation	9	8	7	8	6
Number of Facilities below Expectation	9	8	9	6	7
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	57.1%	46.2%

The developer provided additional documentation stating that the mean score is 49.2, the standard deviation is 17.6, the max is 83.0 and the minimum is 10.0.

Disparities

The developer provides a chart breaking down performance on a case level by gender, ethnicity, payor source, and CMS region. The case level information shows variation and trends for gender, race, payer source, and region for the motor

measure for the years 2007 to 2011. Information is not provided on whether the differences are statistically significant, however, the data to provide information on factors for consideration in assessing variation and impact on various populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. Evidence to Support Measure Focus

Comments:

**There is ample evidence provided that the outcome matters, and is in fact the aim of a LTAC stay. From the flow diagram, it is clear that the provider has influence over the measured outcome, by intensifying or targeting treatment to improve function.

**There is evidence that provision of rehabilitation improves outcomes, and that the FIM (which is the instrument that provides the data elements for the measure) can be used to track this. There is little evidence provided for the actual measure.

**The measure developers have demonstrated this new measure is related to the larger FIM instrument. The evidence relates directly to the outcome being measured, and change in mobility score should relate to functional outcome, such as return to community living. Although not specifically shown by developers, time spent in therapies probably relates to improvement in FIM change.

1b. Performance Gap

Comments:

**There is some case level data supplied that show variation by gender, payer, region most strikingly, but hard to know if this significant. Is there data for change within specific disease groups? There is also some data provided about how many facilities were below or above expectation, with 50% at expectation. That presumably means there is a double digit percentage who could improve.

**The data provided does indicate variation in performance - however the number of LTAC facilities on which this is based is very small. I'm unclear if this indicates the need for a national performance measure, and there is no indication of variation in performance across population subgroups.

**The developers have provided evidence of variability that is regional and payer mix related. The payer source variability may relate to the patient's age, as Medicare was the lowest. They did not report specific statistics looking at the age of patient. The regional variability is hard to explain, but again perhaps is age related, so data related to age and this measure might be useful.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Functional change assessment tool, OASIS

Specifications:

- This is a facility level measure.
- The measure result is a ratio of observed/expected facility average:

- Average change in rasch derived mobility functional score from admission to discharge at the facility level for LTAC patients, over Facility adjusted expected change in rasch derived values, adjusted at the Case Mix Group level, based on impairment type, admission functional status, and age.
- Average is calculated as (sum of change at the patient level/total number of patients).

- The calculation algorithm is included.
- Patients under age 18 and patients who died in the LTAC are excluded.
- A data dictionary is included.
- The measure is stratified by risk category.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Validity/reliability of FIM is documented
- This measure uses a subset of the FIM, so a Rasch analysis was conducted to test:
  - the psychometric properties of the subset of 4 items within the three venues of post-acute care, IRFs, LTACs, and SNFs.
  - The measure reliability at both the person and item level
  - Rasch analysis was used to determine the fit of each item within the measure (4 items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) through infit and outfit statistics and item specific correlations.
- Internal consistency of the critical data elements was demonstrated with Cronbach's alpha
- Reliability must also be demonstrated for the computed performance score (clarification of criteria established by the CSAC in 2016) – the developer has not yet provided this information but us striving to do so prior to the in-person meeting. The developer was provided the following guidance from NQF: *We still do not quite see how the pattern analysis you have provided demonstrates that one can distinguish performance between facilities (perhaps you can explain this a little more?). Note that showing the item-level information is not helpful in demonstrating score-level reliability, as we are interested in the overall performance score, not the item scores. Some folks use the split-half method and calculate an intra-class correlation. To do this analysis, they would randomly assign half of a facility's patients to one dataset and half to another, then do this for all the facilities in their sample. They would then calculate the facility average functional score (for each facility), then calculate the ICC across the facilities.* UDSMR has indicated they are working to fulfill these data needs.

Results of reliability testing

- The developer reports results demonstrating reliability for the subset of the FIM items: the person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.82 to 0.90. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).
- See note above that facility performance score level data is forthcoming from the developer.

Guidance from the Reliability Algorithm
Precise specifications – yes (box 1) -> empirical testing of data elements (box 2) -> TBD

Note: The measure worksheets will be updated prior to the in-person meeting for consideration of the Reliability criterion. We ask the Committee to complete their measure evaluation surveys for the remaining criteria; and are welcome to add notes on Reliability but also acknowledge the developer is working to provide the additional information NQF staff have requested.

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
- Developers used concurrent validity of the FIM total score (all 18 items) with the FIM mobility score: the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (mobility items).
- Predictive validity of the mobility score was tested to determine if the measure predicts outcomes such as functional change and likelihood of discharge to the community setting.

Validity testing results:
- The developer states that both concurrent and predictive validity were correlated with the FIM total score across all venues (IRFs, LTACs, SNFs). The correlations for LTACs are 0.711 ($p < 0.001$) at admission and 0.764 ($p < 0.001$) at discharge. For predicative validity of functional gain, LTACs scored 0.665 ($p < 0.001$), which is considered acceptable and for discharge disposition the C-statistic is 0.79.
- For SNFs, the r-squared values at admission were 0.512 and at discharge 0.707 for functional gain.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients under age 18 and patients that died in the post-acute care setting were excluded. The developer reports these are both consistent with the literature.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developer states the following risk adjustment method: To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The case mix group classification system groups similarly impaired patients based on functional status at admission or patient severity. This is used for SNFs and IRFs, and the same procedure will be applied to the LTACs. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is risk adjusted using Case Mix Group, using an indirect standardization method.
- Statistical tests were not completed, with a rationale that this is a standard procedure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *No information is provided on risk adjustment for SDS factors. Do you think the measure should include SDS factors in the risk adjustment? Why or why not?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provided additional information in an addendum, including "graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers".

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- 2b7 is not included in the form, but in S.22 the developer states that all variables are required, so there should not be missing data. However, if there is missing data, cases should be excluded.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1): Yes: All potential threats to validity relevant to measure empirically assessed (Box 2): Yes and No (suggest discussing risk adjustment further and missing data – we'd typically want to see percentage of cases excluded to indicate if there is impact on the measure – assuming this information can be provided) →Validity testing conducted for computed performance measure score (Box 6): Yes → Method described appropriate (Box 7): Yes →Rating on certainty and confidence that performance measures cores are a valid indicator of quality: Moderate (Rationale: instrument has been demonstrated as valid, testing is appropriate, limited information provided on missing data and risk adjustment)

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**Specifications clear. Case mix adjustment at group level. Algorithms and data dictionary included. No concerns about whether the measure can be consistently implemented.

**The items that make up the measure are taken from the FIM, and how this is calculated is clear. I would have concerns about the likelihood of implementation, given that LTACs already routinely collect similar data using the OASIS. Also - I'm not sure that the measure is appropriate for all patients who are admitted to LTACs.

2a2. Reliability Testing

Comments:

**Deferred to in patient meeting pending testing by developer

**The reliability measures they provide would suggest that there is no consistency in the measure across facilities (am I interpreting this correctly?) and that they say that this is a good thing? I would be concerned about the reliability of the measure.

**The developers showed organization level data.

2b.2 Validity Testing

Comments:

**measure level testing took place. Can't comment on reliability test - see above. Empirically tested at measure level. Agree that improvement in a mobility score indicate that the outcome was met - not sure if show an improved score was about good care and not natural improvement over time.

**The measure does appear to have predictive validity.

**Validity is inferred from these previously used items in the FIM instrument.

2b3.-2b7. Test Related to Potential Threats

Comments:

**Exclusions might have been appropriate, but there is missing data on how many cases were excluded.

Developer says method of case mix adjustment is standard. I am not familiar enough with alternative methods of risk adjustment to know if this is reasonable or not.

Developers included evidence from the literature that the score change does represent a meaningful clinical difference.

**I wasn't clear about the risk adjustment methodology.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are collected during care delivery and are available electronically.
- Commercial use requires a license agreement and has a fee. The developer reports the following:
    - The Functional Change: Change in Motor Score form (this form includes the items for the mobility measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per commercial sale."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**All the the data elements are routinely collected and used to assess efficacy of care as part of a larger instrument in common use. Non concerns that the data collection, particularly if in electronic form, could not be operationalized. Can this be used for free by facilities, not just individual providers?

**The data elements for the FIM are not currently routinely collected by LTACs. At present only a very few use the FIM routinely - and as highlighted above I would be worried about data burden if it were introduced into routine practice, as the LTACs already collect similar data using OASIS, and will eventually be collecting similar data using the CARE tool.

**In the IRF setting, all providers (nursing, therapists, physicians) are trained in the use of the FIM data instrument. In the LTACH setting, there are nurses, nursing aides, and therapists. Training of all staff and achieving better inter-rater reliability will be concerns for the LTACH administrators.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

- The measure is currently used for internal reporting and national benchmarking by LTACs who subscribe to the UDSMR software/outcomes reporting.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting is planned but no details are provided.

Improvement results

- New measure – not available. While a new measure to NQF, the developer does provide trending data for the rasch derived scores from 2007-2011:

Year	2007	2008	2009	2010	2011
Mobility Change Average (Rasch)	18.1	18.8	19.0	18.2	19.8
Case Count	5807	5303	4996	4861	4598
Number of Facilites at or above Expectation	9	8	7	8	6
Number of Facilities below Expectation	9	8	9	6	7
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	57.1%	46.2%

Unexpected findings (positive or negative) during implementation

- None reported

Potential harms

- The developer states that no potential harms were identified since previously collected data was used.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4 Usability and Use

Comments:

**Not publicly reported, but nonspecific plans to do so. FIM reported at facility level for facilities that subscribe to UDSMR. Performance results can be used to incentive well performing facilities, and highlight deficiencies at poorly performing ones. Public reporting is a powerful motivator as well. Would worry that if the case mix adjustment isn't fair, that facilities would preferentially accept patients with a better prognosis or less intensive needs, to the detriment of patient with less potential or greater needs.

**There appear to be only 13 LTACs using the FIM at present (a very small number of the national total). They are using the data it provides for internal purposes and national benchmarking. I have an overall concern that the documentation provides support for the FIM (the data collection tool), rather than the measure.

**Not clear at this point how the data will be publicly reported. To the extent that change in this measure is strongly correlated with return to home discharge destination, or less burden of care for the caregiver, then the performance results on this measure may improve quality of healthcare.

Criterion 5: Related and Competing Measures

Related or competing measures

None listed

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Title: Functional Change: Change in Mobility Score for Long Term Acute Care Facilities
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 3/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus
The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:
Outcome

☒ Health outcome: Functional Status

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

☐ Process: Click here to name the process

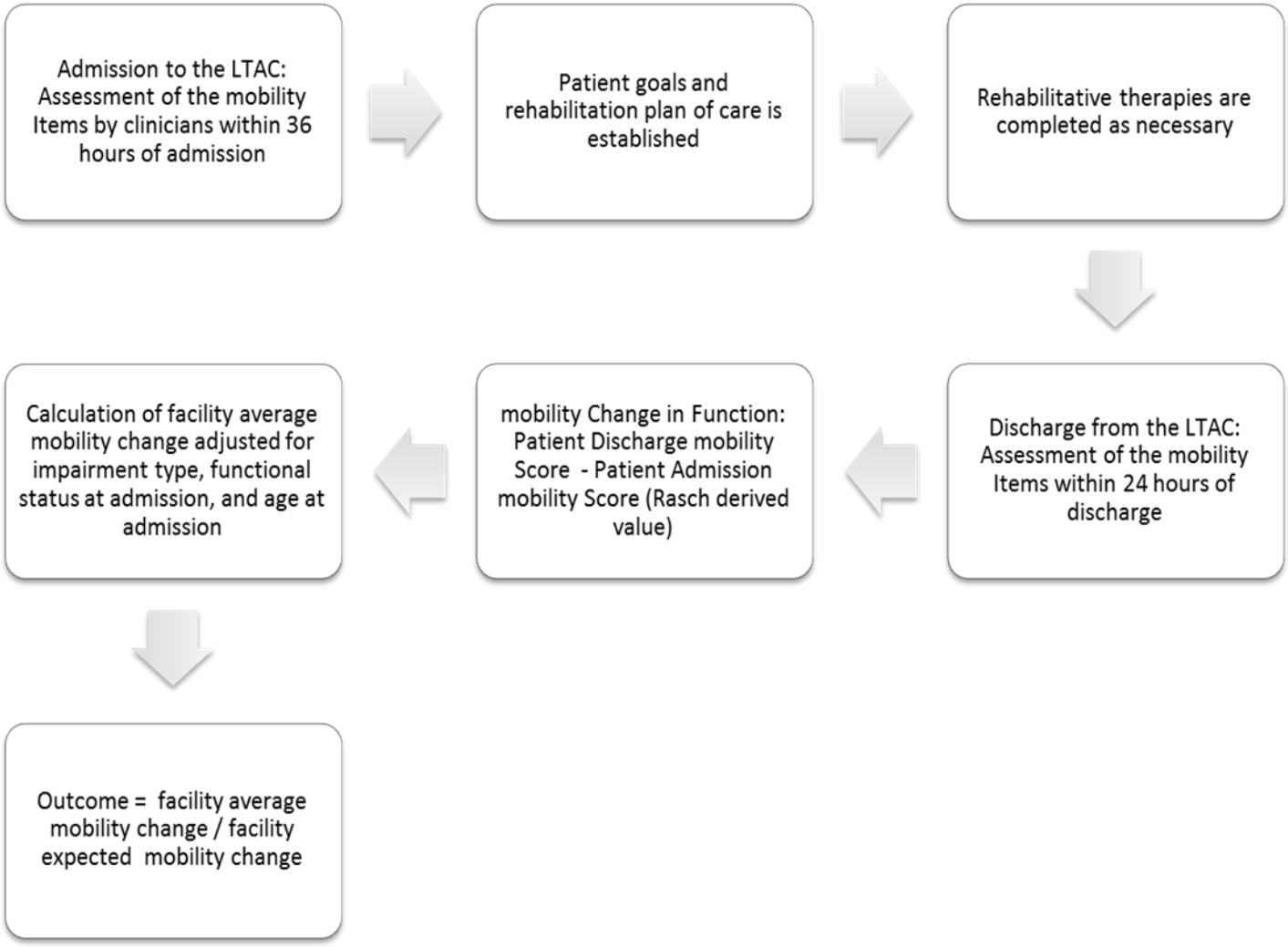
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Long Term Acute Care Hospitals (LTACs) are one part of a multi-level post-acute care continuum. The primary aim of rehabilitation at LTACs is restore function, increase functional independence, and ideally, to discharge the patient back to the community setting or residence prior to the patient's acute admission and/or LTAC stay. While the FIM® ("FIM") instrument is presently embedded in the IRF-PAI, which is the instrument that is presently used in inpatient rehabilitation facilities to assess the patient's level of functional status at admission and at discharge, there are LTACs in the United States that are currently collecting FIM data. It should not be difficult to complete the functional change form for patients seen at LTACs. To date the mobility measure has not been reported on as a stand-alone measure. However, the items of the mobility measure have been extensively used for over twenty five years as a component of the larger 18-item FIM instrument. The mobility measure is intended to be administered within 24 hours of the patient's admission to the IRF and again at patient discharge. Interim assessments can be performed for case management purposes (goal setting or altering the therapy) but are not required. The items that comprise the mobility measure are as follows: The items that comprise the mobility measure are as follows: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. All items are rated by trained clinicians. Below is a flow chart depicting the current methodology for patient assessment in an IRF, which would be the same procedure for LTAC patients:



UDSMR has been a data repository for the items in our proposed measure among LTAC patients for over 20 years. Therefore, data is already available on the measure. Below is a data table displaying aggregate trends for the mobility measure for the years 2007 to 2011 for LTAC patients:

Year	2007	2008	2009	2010	2011
Mobility Change Average (Rasch)	18.1	18.8	19.0	18.2	19.8
Case Count	5807	5303	4996	4861	4598
Number of Facilites at or above Expectation	9	8	7	8	6
Number of Facilities below Expectation	9	8	9	6	7
Percent of Facilities at or above Expectation	50.0%	50.0%	43.8%	57.1%	46.2%

In addition, data are available related to the measure and disparities. Below is a table displaying trends for gender, race, payer source, and region for the mobility measure for the years 2007 to 2011:

Outcomes by group (Gender, Ethnicity, Payer Source, and CMS Region)	2007		2008		2009		2010		2011	
	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)	Case Count	Mobility Change Average (Rasch)
Gender										
Male	3,126	18.8	2,897	19.6	2,724	19.9	2,641	19.1	2,493	20.4
Female	2,676	17.3	2,398	17.8	2,267	18.0	2,215	17.2	2,101	19.2
Ethnicity										
White	4,653	18.4	4,346	18.9	3,895	19.1	3,606	18.0	3,508	19.8
Black	636	18.1	547	19.1	538	19.6	463	17.4	379	19.3
Hispanic	62	19.0	61	23.2	56	19.3	81	18.1	47	17.5
Other Ethnicity	456	14.9	349	16.5	507	17.7	711	20.1	664	20.4
Payer Source										
Medicare	3,444	15.9	3,075	16.1	2,264	15.9	2,222	15.1	2,342	16.8
Medicaid	366	22.6	337	20.2	321	20.3	246	20.7	225	23.3
Commercial	679	19.6	641	21.8	657	20.2	631	19.4	535	22.3
Blue Cross	588	21.7	514	23.0	476	23.8	444	23.0	414	25.7
Other Payer	730	21.7	736	23.7	1,278	21.9	1,318	20.9	1,082	22.2
CMS Region										
P01 (VT, NH, ME, MA, RI, CT)	1,947	20.7	1,953	20.5	2,236	20.1	2,474	18.9	2,622	20.9
P02 (NY, NJ, PR)	221	18.0	0	-	0	-	0	-	0	-
P03 (PA, WV, VA, DE, MD, DC)	436	21.2	364	21.2	358	19.7	419	19.1	369	18.1
P04 (KY, TN, NC, SC, MS, AL, GA, FL)	670	12.7	676	13.4	624	15.6	481	16.7	346	17.1
P05 (MN, WI, IL, IN, MI, OH)	1,774	16.3	1,727	17.5	1,251	17.6	1,043	15.3	765	16.5
P06 (NM, OK, AR, LA, TX)	494	19.0	355	23.3	277	24.7	275	24.5	284	22.8
P07 (NE, IA, KS, MO)	265	17.9	228	18.8	250	17.9	169	19.2	212	22.0
P08 (MT, ND, SD, WY, UT, CO)	0	-	0	-	0	-	0	-	0	-
P09 (CA, NV, AZ, HI)	0	-	0	-	0	-	0	-	0	-
P10 (WA, OR, ID, AK)	0	-	0	-	0	-	0	-	0	-

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

As previously stated, the mobility measure is a new measure and has not been used as a stand-alone tool. However all of the items within the measure are included in a larger instrument (the FIM instrument) which has been widely used and extensively published upon. For these reasons, much of the rationale, feasibility, usability and validity of the mobility measure is referenced to the larger FIM instrument, which is, in essence, the foundation. The validity and utility of the FIM instrument has been demonstrated in hundreds of peer-reviewed journal articles (see bibliography in Appendix). The following are specific to Long Term Acute Care Hospitals:

1. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

2. deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

3. Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE
1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE
1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE
1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE
1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE
1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE
If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

A comprehensive review of the existing, published literature was performed using PubMed and other scholarly search engines. A complete bibliography is maintained by UDSMR for all journal articles using the FIM instrument both nationally and internationally. The bibliography is attached in the Appendix.

1a.8.2. Provide the citation and summary for each piece of evidence.

Abbreviate citations and summaries, along selected articles are discussed below. See Appendix for expanded citations.

Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Archives of physical medicine and rehabilitation.* 2006;87(1):32-39.

OBJECTIVE: To define the minimal clinically important difference (MCID) for the FIM instrument in patients poststroke. DESIGN: Prospective case series discharged over a 9-month period. SETTING: Long-term acute care hospital. PARTICIPANTS: Patients with stroke (N=113). INTERVENTIONS: Not applicable. MAIN OUTCOME MEASURES: Admission, discharge, and change scores were calculated for the total FIM, motor FIM, and cognitive FIM. Assessments of clinical change were rated at discharge on a 15-point (-7 to +7) Likert scale by attending physicians, with MCID defined at a cutoff score of 3. The FIM change scores associated with MCID were identified from receiver operating characteristic curves. Bayesian analysis was used to determine the probability of individual patients achieving MCID. RESULTS: FIM change scores associated with MCID were 22, 17, and 3 for the total FIM, motor FIM, and cognitive FIM, respectively. The accuracy of the MCID was greater when subjects were categorized based on admission FIM scores than when considering the sample as a whole. Larger FIM change scores were related to MCID in subjects with lower admission FIM scores. CONCLUSIONS: These findings will assist in the interpretation of FIM change scores relative to physicians' assessments of important clinical change.

deGuise E, leBlanc J, Feyz M, et al. Long-term outcome after severe traumatic brain injury: the McGill interdisciplinary prospective study. *The Journal of head trauma rehabilitation.* 2008;23(5):294-303.

OBJECTIVE: To obtain a comprehensive understanding of long-term outcome after severe traumatic brain injury (sTBI). PARTICIPANTS: Forty-six patients with sTBI. DESIGN: Comparison of interdisciplinary evaluation results at discharge from acute care and at 2 to 5 year follow-up. MAIN MEASURES: Extended Glasgow Outcome Scale, the FIM instrument, and the Neurobehavioral Rating Scale-Revised. RESULTS: Significant improvement was observed on the FIM instrument, the Extended Glasgow Outcome Scale, and on 3 factors of the Neurobehavioral Rating Scale-Revised. These measures at discharge were significant predictors of outcome. CONCLUSION: Patients with sTBI 2 to 5 years postinjury showed relatively good physical and functional outcome but poorer cognitive and emotional outcome.

Gray DS, Burnham RS. Preliminary outcome analysis of a long-term rehabilitation program for severe acquired brain injury. *Archives of physical medicine and rehabilitation.* 2000;81(11):1447-1456.

OBJECTIVES: To describe the general characteristics and functional outcomes of individuals treated in a publicly funded, long-term, acquired brain injury rehabilitation program and investigate variables affecting functional outcomes in this patient population. DESIGN: Retrospective database review of demographic, descriptive, and functional outcome assessment data. SETTING: Publicly funded, comprehensive, multidisciplinary, long-term, residential brain injury rehabilitation program in Alberta, Canada (64 beds). PATIENTS: All rehabilitation patients admitted to and discharged from the brain injury program from February 1991 to March 1999 (n = 349). INTERVENTIONS: Multidisciplinary rehabilitation program. MAIN OUTCOME MEASURES: Demographic and descriptive information included sex, age at admission, type and severity of injury, time from injury to long-term program admission, and length of stay (LOS). Functional outcome information included level of care required at admission and discharge, admission and discharge Rappaport disability rating scale scores, and admission and discharge FIM instrument and Functional Assessment Measure scores for a subset of patients. RESULTS: Fifty-nine percent of the subjects had severe traumatic brain injuries (TBI) and 41% had severe nontraumatic brain injuries (NTBI) of various causes. Mean age at admission was older and LOS was longer for NTBI compared with TBI; there were no other differences between the groups in demographic or descriptive measures. The TBI group had significantly lower admission motor subscale scores than the NTBI group, but the groups did not differ on cognitive scores. All functional assessment measures showed statistically significant improvement from admission to discharge, and 85.6% of patients were discharged to community living after a mean LOS of 359.5 days. Functional

status at admission, age at admission, length of time between injury and admission, and LOS in the rehabilitation program significantly correlated with functional improvement. CONCLUSIONS: Patients with severe TBI and NTBI who were not candidates for other more conventional forms of rehabilitation showed significant improvement in functional outcomes after extended program admissions. Consideration was also given to the potential insensitivity of commonly used outcome assessment measures in this population.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2778
De.2. **Measure Title:** Functional Change: Change in Mobility Score for Long Term Acute Care Facilities
Co.1.1. **Measure Steward:** Uniform Data System for Medical Rehabilitation, a
De.3. **Brief Description of Measure:** Change in rasch derived values of mobility function from admission to discharge among adult LTAC patients aged 18 years and older who were discharged alive. The time frame for the measure is 12 months. The measure includes the following 4 mobility items:Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.
1b.1. **Developer Rationale:** The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. There are LTACs that are currently collect data on the items in the proposed measure for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community. The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care.
We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care after their LTAC stay.

S.4. **Numerator Statement:** Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.
S.7. **Denominator Statement:** Facility adjusted adjusted expected change in rasch derived values, adjusted at the Case Mix Group level.
S.10. **Denominator Exclusions:** Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Measure_Evaluation_Mobility_LTAC-635950314051745274.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The current mandated quality measures for Long Term Acute Care facilities do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or commercial payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. The primary aim of rehabilitation is to increase function to return the patient to living in the community or to another less intensive venue of care. Yet the current measures don't adequately capture function or functional improvement. There are LTACs that are currently collect data on the items in the proposed measure for outcomes purposes; therefore, it should not be difficult for all LTACs to collect this additional information. The change in mobility measure has demonstrated both reliability and validity as results indicated a high overall internal consistency, the ability to capture significant functional gains during rehabilitation, has high discriminative capabilities for rehabilitation patients, and predictive of change in mobility function outcomes and likelihood of patient discharge from inpatient rehabilitation to the community. The current mandated quality measures for LTACs do not adequately address the rehabilitative objectives or functional status of patients. The measures do not allow facilities to substantiate the quality of their restorative care program to CMS or payers. The emphasis on restoration or maintenance of function affected by the patient´s illness or injury is paramount in the episode of care. We feel it is imperative that any quality indicators used for the PAC setting take into account the overriding goal of rehabilitation outcomes, which is to restore and improve function and increase functional independence among individuals receiving rehabilitation, and by doing so allowing the patient the ability to return to a community setting upon discharge or other less intensive venue of care after their LTAC stay.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see measure evaluation form.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see measure evaluation form.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Functional Status, Health and Functional Status, Health and Functional Status : Development/Wellness, Health and Functional Status : Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_Submission_Mobility-635749871757956568.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Average change in rasch derived mobility functional score (Items Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) from admission to discharge at the facility level. Average is calculated as (sum of change at the patient level/total number of patients). Cases aged less than 18 years at admission to the facility or patients who died within the facility are excluded.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. The numerator is the average change in rasch derived mobility functional score from admission to discharge for each patient at the facility level, including items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs. Average is calculated as: (sum of change at the patient level for all items (Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) / total number of patients).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Facility adjusted adjusted expected change in rasch derived values, adjusted at the Case Mix Group level.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Dual eligible beneficiaries, Populations at Risk : Individuals with multiple chronic conditions, Populations at Risk : Veterans, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is all LTAC patients, at least 18 years old, who did not die in the LTAC. Impairment type is defined as the primary medical reason for the LTAC stay (such as stroke, joint replacement, brain injury, etc.). Admission functional status is the expected value of the average of the sum 4 items (Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs) at the facility level. Age is the age of the patient at the time of admission to the LTAC. The denominator is meant to reflect the expected Mobility functional change score at the facility, if the facility had the same distribution of CMGs (based on impairment type, functional status at admission,and age at admission). This adjustment procedure is an indirect standardization procedure (observed facility average/expected
facility average).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Excluded in the measure are patients who died in the LTAC or patients less than 18 years old.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Living at discharge and age at admission are collected through OASIS

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
See definition of the CMGs in the excel file provided.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
This adjustment procedure is an indirect standarization procedure (observed facility average/expected facility average). The numerator is the facility´s average mobility functional change score. The denominator is meant to reflect the expected Mobility functional change score at the facility, if the facility had the same distribution of CMGs(impairment, functional status at admission, and age at admission).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify all patients during the assessment time frame (12 months).
2. Exclude any patients who died in the LTAC.
3. Exclude any patients who are less than 18 at the time of admission to the LTAC.
3. Calculate the total mobility change score for each of the remaining patients (sum of change at the patient level for all items (Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.)
4. Transform the patient level functional change scores to the rasch derived value (as stated in excel file).
5. Calculate the average rasch derived mobility change score at the facility level.
6. Using national data and previously described adjustment procedure, calculate the facility´s expected rasch derived average mobility change score for the time frame (12 months).
7. Calculate the ratio outcome by taking the observed facility average mobility change score/facility´s national expected mobility change score.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is not based on a sample, but rather is meant for all patients minus the exclusion criteria.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This is not a survey/patient reported measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There should not be missing data for this measure as all variables would be required, however, should data be missing, those cases will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Functional Change Form, as seen in the appendix.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Post Acute/Long Term Care Facility : Long Term Acute Care Hospital
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_Testing_Mobility_LTAC.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Functional Change: Change in Mobility Score for Long Term Acute Care Facilities
Date of Submission: 3/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation

counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

FIM® ("FIM") instrument data from inpatient rehabilitation facilities (IRFs), long term acute care (LTACs), and skilled nursing facilities (SNFs) from the Uniform Data System for Medical Rehabilitation (UDSMR). The UDSMR, a not-for-profit organization affiliated with the UB Foundation Activities, Inc. at the State University of New York at Buffalo, maintains the largest non-governmental database for medical rehabilitation outcomes.

1.3. What are the dates of the data used in testing? Years 2010-2012 were used for the mobility measure development (reliability and validity testing, Rasch modeling for establishing psychometric properties of the measure). Years 2002-2013 were used in examining the data trends over time using the mobility measure and patient outcomes of long term acute care facilities.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan

☐ other: Click here to describe	☒ other: patient level/aggregate

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

All three post-acute care hospital based venues are included, inpatient rehabilitation facilities (n = 746), long term care hospitals (n = 6), and skilled nursing facilities (n = 174). All facilities subscribed to UDSMR for outcomes reporting and severity adjusted benchmark analyses.

Of the 746 inpatient rehabilitation facilities included, 571 (76.5%) were units within an acute care hospital and 175 (23.5%) were free-standing IRFs. Every state in the U.S. was represented among the 746 facilities.

Of the 6 long term acute care hospitals (LTCHs), three were in Massachusetts, one was in Missouri, one was in Michigan, and one was in South Carolina.

Of the 174 skilled nursing facilities (SNFs), 141 (84.4%) were free-standing facilities, and 26 (15.6%) were located in an acute care hospital. Twenty-three of the 50 United States were represented.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We used a random sample of 11,525 patients for all three venues so that one venue was not over sampled in the analysis (to avoid overrepresentation of IRFs and underrepresentation of SNFs and LTCHs) and comparable case counts were included from each venue of care, IRFs (n = 3,619), LTACs (n = 3,922), and SNFs (n = 3,984). Below is a table displaying the demographic distribution.

	Total	IRFs	LTACs	SNFs
	n = 11,525	n = 3,619	n = 3,922	n = 3,984
Age, mean (SD)	70.2 (15.5)	69.2 (15.4)	76.1 (11.7)	65.2 (16.8)
Age Groups, count (%)				
44 years old or less	748 (6.5)	250 (6.9)	447 (11.4)	51 (1.3)
45 to 65 years old	2,782 (24.1)	961 (26.6)	1,229 (31.3)	592 (14.9)
65 to 74 years old	2,733 (23.7)	858 (23.7)	950 (24.2)	925 (23.2)
75 years and older	5,262 (45.7)	1,550 (42.8)	1,296 (33.0)	2,416 (60.6)
Rehabilitation Impairment Category, count (%)				
Stroke	1,547 (13.4)	784 (21.7)	553 (14.1)	210 (5.3)
Traumatic Brain Dysfunction	395 (3.4)	146 (4)	224 (5.7)	25 (0.6)
Non-traumatic Brain Dysfunction	344 (3)	195 (5.4)	103 (2.6)	46 (1.2)
Traumatic Spinal Cord Dysfunction	129 (1.1)	43 (1.2)	82 (2.1)	4 (0.1)
Non-traumatic Spinal Cord Dysfunction	219 (1.9)	152 (4.2)	54 (1.4)	13 (0.3)
Neurological Conditions	536 (4.7)	396 (10.9)	72 (1.8)	68 (1.7)
Lower Extremity Fracture	736 (6.4)	381 (10.5)	27 (0.7)	328 (8.2)
Lower Extremity Joint Replacement	1,084 (9.4)	363 (10)	46 (1.2)	675 (16.9)
Other Orthopaedic Conditions	670 (5.8)	222 (6.1)	92 (2.3)	356 (8.9)
Lower Extremity Amputation	180 (1.6)	111 (3.1)	40 (1)	29 (0.7)
Other Amputation	20 (0.2)	1 (0)	8 (0.2)	11 (0.3)
Osteoarthritis	39 (0.3)	9 (0.2)	3 (0.1)	27 (0.7)
Rheumatoid and Other Arthritis	50 (0.4)	25 (0.7)	8 (0.2)	17 (0.4)
Cardiac Conditions	601 (5.2)	147 (4.1)	124 (3.2)	330 (8.3)
Pulmonary Disorders	429 (3.7)	47 (1.3)	179 (4.6)	203 (5.1)
Pain Syndromes	114 (1)	29 (0.8)	18 (0.5)	67 (1.7)
Major Multiple Trauma w_o TBI, SCI	182 (1.6)	105 (2.9)	46 (1.2)	31 (0.8)
Major Multiple Trauma with TBI, SCI	110 (1)	58 (1.6)	49 (1.2)	3 (0.1)
Guillain-Barré Syndrome	28 (0.2)	15 (0.4)	12 (0.3)	1 (0)
Miscellaneous	4,102 (35.6)	384 (10.6)	2,181 (55.6)	1537 (38.6)
Burns	10 (0.1)	6 (0.2)	1 (0)	3 (0.1)
Gender, count (%)				
Missing	847 (7.3)	2 (0.1)	5 (0.1)	840 (21.1)
Male	4,991 (43.3)	1,663 (46.0)	2,195 (56)	1,133 (28.4)
Female	5,687 (49.3)	1,954 (54.0)	1,722 (43.9)	2,011 (50.5)

While the above data is displayed at the case level, facility level outcomes and comparisons at the facility level can be supplied if required.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The validity and reliability of the FIM instrument (the tool used for this measure) is well documented, including inter – and intra-rater reliability[1-7]. The measure proposed, however, uses only a subset of the FIM instrument items. Therefore, Rasch analysis was conducted to test the psychometric properties of the subset of 4 items within the three venues of post-acute care, IRFs, LTACs, and SNFs. It is understood the proposed measure is intended for long term acute care facilities. However, we are aware that there has been a number of policy reports indicating the importance for a measure to be capable of use in all inpatient post-acute care venues. Additionally, it is well-recognized that policies such as site neutral payments and bundle payments have been proposed. Our mobility measure is appropriate for use in multiple post-acute care venues, which is a strength of the measure as it is advantageous to collect the exact same items which measure the same construct using the same risk adjustment methodology in all inpatient post-acute care to be able to compare outcomes, quality and value of care by setting and among patients that may have used several post-acute care venues for rehabilitation.

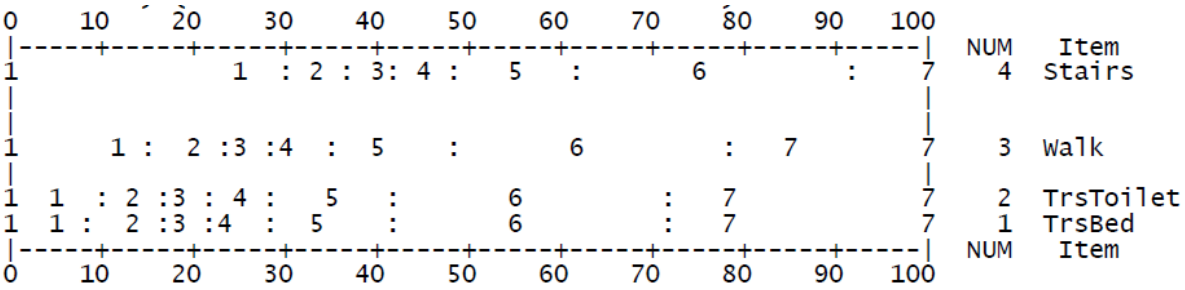
Rasch analysis was used to determine the measure reliability at both the person and item level, as well as internal consistency through the use of Cronbach's alpha. Rasch analysis was also used to determine the fit of each item within the measure (4 items: Transfer Bed/Chair/Wheelchair, Transfer Toilet, Locomotion and Stairs.) through infit and outfit statistics and item specific correlations. We used Winsteps 3.73 for the analysis.

In addition, Rasch analysis allows for the conversion of ordinal-level data into interval-level data. Ordinal measures do not inherently act as interval measures, where the difference between one score is equidistant compared to the difference between another two scores, i.e. the difference between a 15 and a 16 in our measure may not reflect the same difference between a 56 and a 57, in terms of difficulty. If the data fit the Rasch model, a result of the analysis is the conversion of the raw ordinal scores to a Rasch derived interval score. This allows for a more precise estimation of differences in functional status both between patients and across facilities.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The person-reliability correlation was 0.89. The Cronbach Alpha reliability statistic was 0.92. Item correlations within the measure ranged from 0.82 to 0.90. In addition, the infit and outfit statistics were acceptable for all items (less than 2.0).

For the conversion of the ordinal level measure to an interval measure, we set the Rasch scale at 0 – 100 with a high value indicating more independence. The following figure displays the "ruler" or interval transformation scores for each item in the measure.



The ruler shows that the easiest item is Transfers: Bed/Chair/Wheelchair, and the hardest Stairs and that the distances between a level 1 and 2 and 5, 6 and 7 are greater than the distances between the remaining levels of each item. When calculated at the total level, the following table displays the Rasch-transformed values at each possible raw value.

TABLE OF MEASURES ON TEST OF 4 Item

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
4	.00E	12.48	13	26.23	3.76	22	50.27	5.85
5	8.05	6.65	14	28.41	3.94	23	55.99	6.63
6	12.45	4.65	15	30.76	4.06	24	62.97	7.09
7	15.07	3.91	16	33.17	4.04	25	70.32	7.08
8	17.10	3.59	17	35.50	3.95	26	77.95	7.52
9	18.91	3.46	18	37.76	3.93	27	87.92	9.24
10	20.65	3.45	19	40.08	4.07	28	100.00E	13.82
11	22.41	3.50	20	42.69	4.42			
12	24.25	3.61	21	45.94	5.04			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As indicated previously, the reliability of the FIM instrument is well known. The results of the analysis for the measure proposed show the reliability holds even when looking at a subset of FIM instrument items.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Since the validity of the 18-item FIM instrument has been well established, we examined the concurrent validity of the mobility measure with the FIM total score, both at admission and discharge. In particular, we used the FIM total score from all 18 items as our gold standard measure in which to test our new mobility measure against. The two tests of validity we used were the Pearson correlation coefficient and linear regression to calculate an r-squared which represents the percent of variance of the dependent variable (FIM total) explained by the independent variable (mobility items). In this instance we examined the admission and discharge values separately.

We assessed the predictive validity of the mobility measure to determine if the measure predicts outcomes such as: functional change (total functional gain as assessed with the 18 item FIM instrument (the gold standard)), and likelihood of discharge to the community setting. Linear regression was used to determine functional change, whereas the change in mobility was the independent variable, the r-squared value (proportion of change accounted for) and the Pearson correlation coefficient was examined. For discharge disposition, logistic regression was used, admission mobility total was the independent variable and the dependent variable was dichotomized as discharge to the community (yes or no). We used the C-statistic derived from the area under the ROC curve to determine the discrimination of the model, or the ability of the model to discriminate between those patients having the outcome of interest or not, as predicted by our measure. In SPSS this is completed by utilizing the patient level probabilities created during the logistic regression in the ROC curve analysis. The C-statistic ranges from 0.5 (no predictive ability) to 1.0 (perfect discrimination).

We completed all testing for the total data set including all venues, and separately by venue of post-acute care. For all analyses, the Rasch derived values for the mobility measure was used. SPSS version 21 was used in the analyses.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Concurrent Validity

Correlations: For all venues, our measure at both admission and discharge was correlated with the FIM total, 0.671 ($p < 0.001$) and 0.768 ($p < 0.001$), respectively. The correlations remained significant within each venue of care; IRFs, 0.605 ($p < 0.001$) and 0.847 ($p < 0.001$); LTACs, 0.711 ($p < 0.001$) and 0.764 ($p < 0.001$); SNFs, 0.659 ($p < 0.001$) and 0.787 ($p < 0.001$).

Linear Regression: For all venues, when comparing our measure at admission and discharge to the respective FIM® totals, the r-square values ranged from respectable for admission FIM total, to high for discharge FIM total, 0.512 and 0.706, respectively. The values remained similar at the venue specific level as well; IRFs, 0.400 and 0.676; LTACs, 0.540 and 0.707; SNFs, 0.454 and 0.707.

Predictive Validity

Functional Gain: For all venues, when comparing gain in our measure to overall FIM gain including all items, the correlation was acceptable, 0.615 ($p < 0.001$). In addition, by venue, the correlations remained acceptable; IRFs, 0.598 ($p < 0.001$); LTACs, 0.665 ($p < 0.001$); SNFs, 0.611 ($p < 0.001$). The linear regression showed acceptable r-squared values as well; all venues, 0.506; IRFs, 0.438; LTACs, 0.559; SNFs, 0.486.

Discharge Disposition – Community: For all venues, the logistic regression analysis shows that the gain in our measure has good predictive ability for discharge setting (community), with a C-statistic of 0.79. By venue, the results are similar; IRFs, 0.78; LTACs, 0.77; SNFs, 0.77.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results show good validity across all analyses. The r-square values were all consistent around 0.5 – 0.6, meaning that the percent of variance explained in the dependent variables by our measure were all more than 50%. Considering we are testing the correlation between 4 items of an 18 item scale, these r-squared values are quite good. In addition, the predictive validity was also high.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients thatdied in the post-acute care setting (an unanticipated outcome) and patient less than age 18, both criteria consistent with published literature examining rehabilitation outcomes.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 1 **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

We used Case Mix Group as our only adjustment variable through an indirect standardization method.

To calculate the facility's adjusted expected change in Rasch derived values, we use indirect standardization which weights national CMG-specific values by facility-specific CMG proportions. CMG-adjustment derives the expected value based on the case mix and severity mix of each facility. The case mix group classification system groups similarly impaired patients based on functional status at admission or patient severity. This is used for SNFs and IRFs, and the same procedure will be applied to the LTACs. Patients within the same CMG are expected to have similar resource utilization needs and similar outcomes. There are three steps to classifying a patient into a CMG at admission:

1. Identify the patient's impairment group code (IGC).
2. Calculate the patient's weighted motor index score, calculated from 12 of the 13 motor FIMinstrument items.
3. Calculate the cognitive FIM total rating and the age at admission. (This step is not required for all CMGs.)

See file uploaded in S.15 for calculations.

2b4.4. What were the statistical results of the analyses used to select risk factors?

No statistical tests were calculated, CMG adjustment is a standard procedure.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

References

1. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. *Archives of physical medicine and rehabilitation.* May 1993;74(5):531-536.

2. Gerrard P, Goldstein R, Divita MA, et al. Validity and Reliability of the FIM(R) Instrument in the Inpatient Burn Rehabilitation Population. *Archives of physical medicine and rehabilitation.* Mar 5 2013.
3. Granger CV, Deutsch A, Russell C, Black T, Ottenbacher KJ. Modifications of the FIM instrument under the inpatient rehabilitation facility prospective payment system. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* Nov 2007;86(11):883-892.
4. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. *Archives of physical medicine and rehabilitation.* Nov 1999;80(11):1471-1476.
5. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil.* 1987;1:6-18.
6. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. *Archives of physical medicine and rehabilitation.* Dec 1996;77(12):1226-1232.
7. Stineman MG, Shea JA, Jette A, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. *Archives of physical medicine and rehabilitation.* Nov 1996;77(11):1101-1108.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
While this is a new measure, the data collection procedure is in place for LTACs utilizing UDSMR software.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The Functional Change: Change in Motor Score form (this form includes the items for the mobility measure) submitted is copyrighted, however, it can be reproduced and distributed, without modification, for internal reporting of performance data or internal auditing that is for non-commercial purposes, e.g. use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Functional Change: Change in Motor Score form for commercial
gain, or incorporation of the Functional Change: Change in Motor Score form into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Functional Change: Change in Motor Score form requires a license agreement between the user and UDSMR. The fees charged for other uses or commercial uses shall be in the range of 0% – 15% per
commercial sale.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) UDSMR www.udsmr.org Quality Improvement (Internal to the specific organization) UDSMR www.udsmr.org

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Currently UDSMR provides both internal reporting and national benchmarking for LTACs who subscribe to the UDSMR software/outcomes reporting. The FIM System® is a an outcomes management program for skilled nursing facilities, subacute facilities, long-term care hospitals, Veterans Administration programs, international rehabilitation hospitals, and other related venues of care. The FIM System® enables providers and programs to document the severity of patient disability and the results of medical rehabilitation and establishes a common measure for the comparison of rehabilitation outcomes.

The FIM System® provides an established means of collecting rehabilitation data in a consistent manner. It allows clinicians to follow changes in the functional status of their patients from the start of rehabilitative care through discharge and follow-up.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
As we used existing data that has already been collected, there were no unintended negative consequences to individuals or populations identified during our testing

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Functional_Change_Appendix-635749878241675737.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Uniform Data System for Medical Rehabilitation, a
Co.2 **Point of Contact:** Paulette, Niewczyk, pniewczyk@udsmr.org, 716-817-7868-
Co.3 **Measure Developer if different from Measure Steward:** Uniform Data System for Medical Rehabilitation, a
Co.4 **Point of Contact:** Margaret, DiVita, mdivita@udsmr.org, 716-817-7800-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 03, 2016
Ad.4 **What is your frequency for review/update of this measure?** Unknown, new measure
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

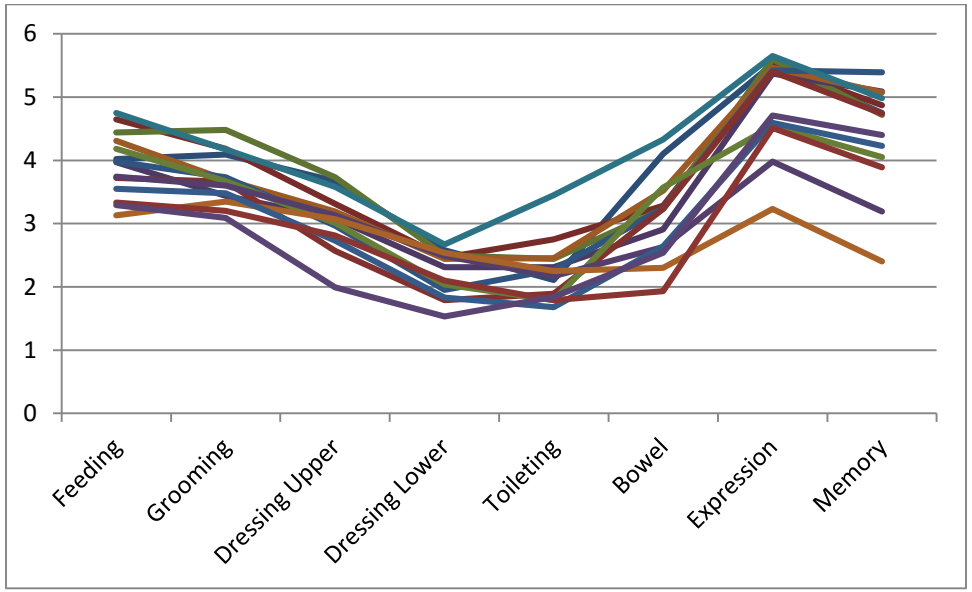
April 28, 2016

Dear NQF, Patient and Family Centered Measures Committee:

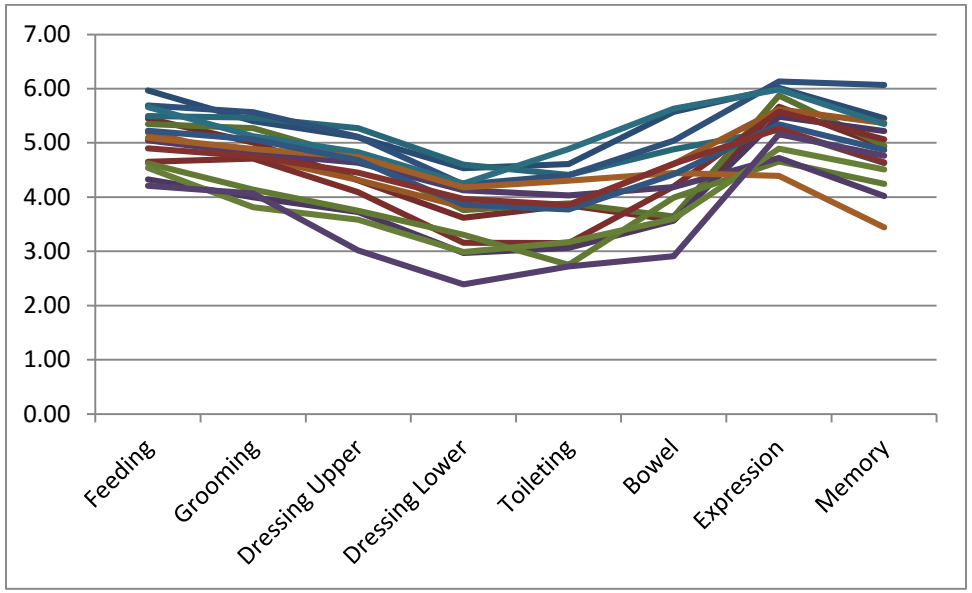
This document is submitted in response to the request by the NQF, Patient and Family Centered Measures Committee for additional information related to the three measures submitted by UDSMR, Change in Function: Self Care Measure for Long Term Acute Care Facilities, Change in Function: Mobility Measure for Long Term Acute Care Facilities and the Change in Function: Motor Measure for Long Term Acute Care Facilities. We have included all of the requested information below, embedded in the subsequent pages of this document.

While the committee requested facility level reliability analyses, and in the past has suggested the Intra-class Correlation Coefficient (ICC), we respectfully maintain that the ICC is not an appropriate statistical test for the type of data maintained in our repository and the very large size of our database. As each of the measures are contained within the larger, FIM Instrument, the inter-rater and intra-rater reliability, validity and psychometric properties has been well established and results have been published in a many peer-reviewed journals; attached is a separate document listing the published references. As an alternative for the ICC analysis request, we provided a rating pattern analyses for each measure, at the item level, for facilities in our database, displayed below. The graphs illustrate that although the values of admission and discharge scores for each item included in our measure may range between facilities, the overall pattern is maintained for the vast majority of facilities, with very few outliers. Each line represents a different facility's average score at each item within the measure. Please note, only data for the self-care and mobility measure are displayed as the motor measure, is simply the combination of the items within the self-care and mobility measures. The graphs illustrate the high consistency in ratings for the items included in all measures.

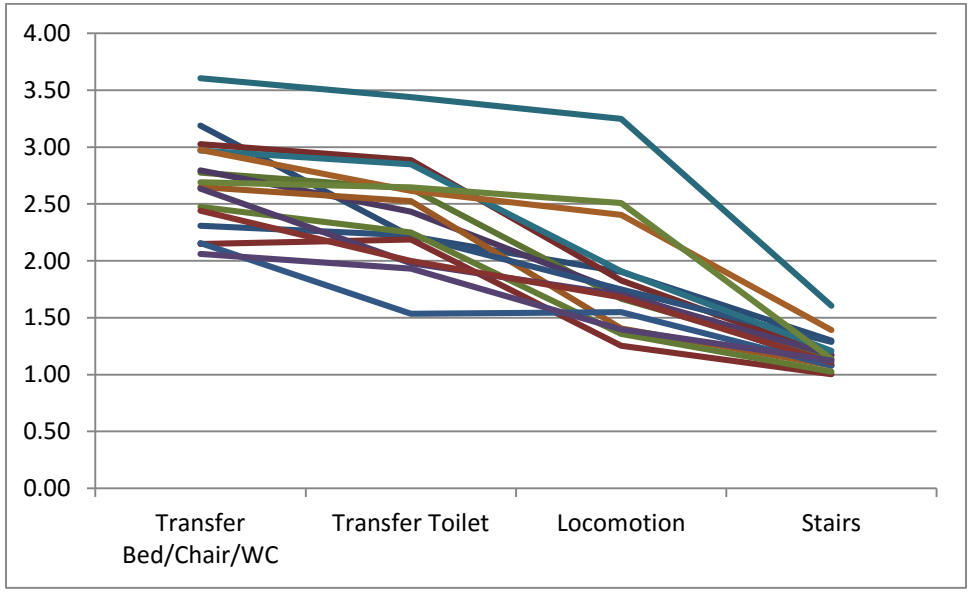
Self-Care Graph: Admission (Year 2009)



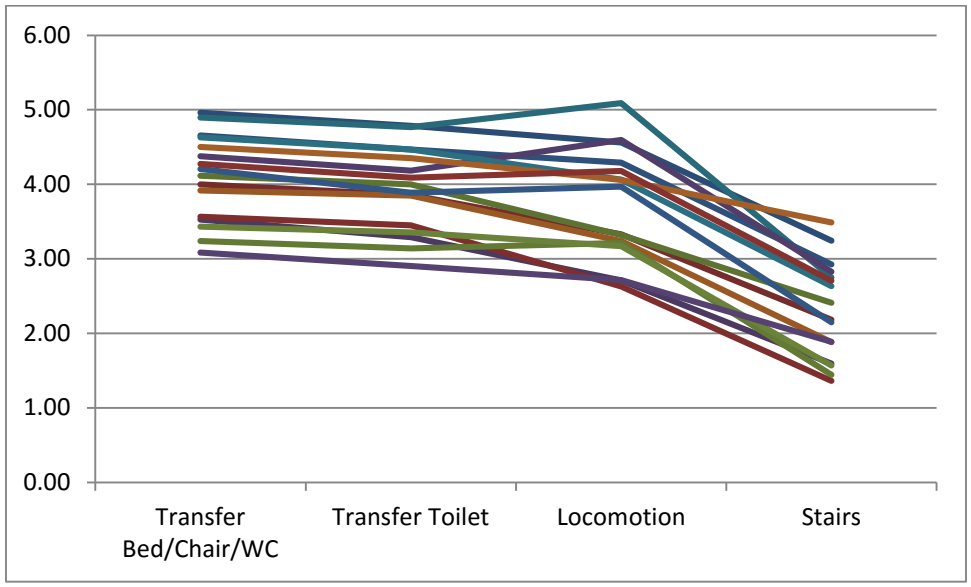
Self-Care Graph Discharge (Year 2009)



Mobility Graph: Admission (Year 2009)



Mobility Graph: Discharge (Year 2009)



Lastly, the mean fit statistics from the rasch analysis for each measure were requested, each are displayed below. Since our measure is meant to be used across the PAC venues of IRFs, SNFs, and LTACs, the rasch analysis was completed using data from all three venues of care, as were the expectations for the measures. Therefore, the following mean fit statistics hold for the LTAC venue of care.

Self-Care Mean Fit Statistics

```
TABLE 3.1 Self Care 8 Items                    ZOU018WS.TXT  Mar 19  9:16 2015
INPUT: 3096 Person  8 Item  REPORTED: 3094 Person  8 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2969 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      36.6       8.0       50.76     3.96       .96    -.1   1.02     .0  |
| S.D.      11.5        .3       13.60     1.46       .71    1.2    .82    1.2  |
| MAX.      55.0       8.0       87.04    10.90      6.32    5.4   8.33    6.2  |
| MIN.       8.0       3.0       11.87     3.00       .05   -3.9    .05   -3.7  |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.60 TRUE SD   12.80  SEPARATION  2.78  Person RELIABILITY  .89  |
|MODEL RMSE   4.22 TRUE SD   12.93  SEPARATION  3.06  Person RELIABILITY  .90  |
| S.E. OF Person MEAN = .25                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     50 Person
  MINIMUM EXTREME SCORE:     75 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      36.2       8.0       50.33     4.59                                |
| S.D.      12.4        .3       16.71     3.40                                |
| MAX.      56.0       8.0      100.06    19.89                                |
| MIN.       8.0       3.0        -.06     3.00       .05   -3.9    .05   -3.7  |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.99 TRUE SD   15.60  SEPARATION  2.61  Person RELIABILITY  .87  |
|MODEL RMSE   5.71 TRUE SD   15.70  SEPARATION  2.75  Person RELIABILITY  .88  |
| S.E. OF Person MEAN = .30                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .93
```

Mobility Mean Fit Statistics

```
                                                    03-448WS
TABLE 3.1 Mobility 4 Items IRF Only                ZOU448WS.TXT  Mar 19  9:38 2015
INPUT: 3096 Person  5 Item  REPORTED: 3088 Person  4 Item  7 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 2558 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      13.8       3.7       31.44    4.51       .94    -.3    .94    -.2 |
| S.D.       6.2        .5       16.49    1.26      1.27    1.4   1.34    1.2 |
| MAX.      27.0       4.0       87.88    9.51      9.90    5.8   9.90    8.5 |
| MIN.       2.0       1.0        8.08    3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   5.45 TRUE SD   15.56  SEPARATION  2.85  Person RELIABILITY  .89 |
|MODEL RMSE   4.68 TRUE SD   15.81  SEPARATION  3.38  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .33                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:     18 Person
  MINIMUM EXTREME SCORE:    512 Person
      LACKING RESPONSES:      8 Person

     SUMMARY OF 3088 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      12.2       3.7       26.70    5.88                                |
| S.D.       6.9        .6       19.75    3.22                                |
| MAX.      28.0       4.0       99.95   13.79                                |
| MIN.       1.0       1.0         .02    3.45       .00   -3.5    .00   -3.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE   7.17 TRUE SD   18.40  SEPARATION  2.57  Person RELIABILITY  .87 |
|MODEL RMSE   6.70 TRUE SD   18.57  SEPARATION  2.77  Person RELIABILITY  .88 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92
```

Motor Mean Fit Statistics

```
TABLE 3.1 All Facilities 12 items                ZOU439WS.TXT  Mar 19  9:43 2015
INPUT: 3096 Person  12 Item  REPORTED: 3094 Person  12 Item  7 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 3013 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      49.2      11.6       45.63     2.83       .99    -.1   1.06     .0 |
| S.D.      17.6        .7       12.31      .98       .67    1.4    .91    1.4 |
| MAX.      83.0      12.0       88.22     9.85      5.13    5.2   9.90    7.7 |
| MIN.      10.0       4.0       10.53     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   3.30 TRUE SD   11.86  SEPARATION  3.59  Person RELIABILITY  .93 |
|MODEL RMSE   2.99 TRUE SD   11.94  SEPARATION  3.99  Person RELIABILITY  .94 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:      7 Person
  MINIMUM EXTREME SCORE:     74 Person
      LACKING RESPONSES:      2 Person

     SUMMARY OF 3094 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      48.4      11.7       44.66     3.21                               |
| S.D.      18.3        .7       14.26     2.51                               |
| MAX.      84.0      12.0      100.06    17.81                               |
| MIN.      10.0       4.0        -.05     2.23       .09   -4.2    .11   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE   4.30 TRUE SD   13.59  SEPARATION  3.16  Person RELIABILITY  .91 |
|MODEL RMSE   4.07 TRUE SD   13.66  SEPARATION  3.36  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .26                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95
```

We appreciate the opportunity to provide the Committee the additional information related to our measures and we welcome any additional questions or clarification needed by the Committee. We thank the NQF and the PFCM Committee for their interest in our measures.

Respectfully,
Paulette M. Niewczyk, MPH, PhD
UDSMR, Director of Research

Margaret DiVita, MS, PhD
UDSMR, Senior Research Analyst

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2958
Measure Title: Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery
Measure Steward: Massachusetts General Hospital
Brief Description of Measure: The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60% or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision.
The target population is adult patients who had a primary hip or knee replacement surgery for treatment of osteoarthritis.
Developer Rationale: Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

Numerator Statement: The numerator is the number of respondents who have an adequate knowledge score (60% or greater) and a clear preference for surgery.
Denominator Statement: The denominator includes the number of surveys of patients who have undergone primary knee or hip replacement surgery for osteoarthritis. Participants who answer at least 3 of the 5 knowledge items and the preference item will be counted in the denominator.
Denominator Exclusions: Respondents who are missing 3 or more knowledge items do not get a total knowledge score and are not able to be assessed for the measure. Similarly, respondents who do not indicate a preferred treatment do not get counted in the denominator.

Measure Type: PRO
Data Source: Patient Reported Data/Survey
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale. In addition to the evidence required for any outcome. The evidence for a Patient-reported outcome-based performance measures (PRO-PM) should demonstrate that the target population values the measured PRO and finds it meaningful.

Summary of evidence:

- The developer states:
  - *The measure is a PRO that reflects the quality of the treatment decision making process. The measure reflects multiple care processes and outcomes such as communication, provision of information, shared decision making, and patient engagement.*
  - *Further, the use of patient decision aids has been associated with increased decision quality. Further, increased decision quality, and having treatments that match patients' preferences, has been associated with reduced utilization of joint replacement surgery and better health outcomes. [Sepucha et al 2011; Sepucha et al 2013; Stacey et al 2014]*
  - *The development of the items included in the measure was conducted with considerable engagement of patients and multidisciplinary group of clinicians. Patients and clinicians rated the importance of the items for assessing informed decision making, and the ones included in the measure not only performed well in psychometric analyses but also were rated highly by patient and clinician stakeholders. [Sepucha 2008] (see item 1c.5)*

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- This is a new measure and the developer provides the following on performance gap:
  - The sample includes patients from three sites and a general population sample from the Boston area. The site that had a formal shared decision making process (SDM site) had a higher rate of informed, patient centered (IPC) surgery than the sites with no formal shared decision making (usual care sites). The association between SDM site and rates of IPC surgery remained significant in multivariate analyses controlling for joint (knee/hip), gender, surgery, and decision making process scores [Sepucha et al 2013].
  - The DECISIONS study was a national random sample of patients surveyed by telephone up to two years after their decision. They asked earlier versions of four of these knowledge items and found that on the whole, patients had considerable knowledge gaps. For the 141 patients who had discussed hip or knee replacement surgery with their health care provider, the total knowledge score was 32.1% (out of 100%. Note that "passing" this measure is >60%)[Fagerlin 2010].
- The above information, in addition to the provided testing information provide evidence of gap in informed, patient-centered decision making in usual care practice sites.

Disparities

The data come from a sample of patients who were surveyed about one year after surgery or after a visit with an orthopedic surgeon. These data suggest that there are differences according to educational status and race but these differences are not statistically significant (the developer suggests this is due to small sample sizes). The differences between age and gender groups were less dramatic and also not statistically significant.

Table: Disparities Data for Knee and Hip Replacement Surgery

	IPC	N	P-value
EDUCATION			
COLLEGE	57.7%	208	.09
< COLLEGE	48.8%	160	

RACE			
NON-HISPANIC WHITE	54.5%	352	.08
OTHER RACES	31.2%	16	
AGE			
<65	52.9%	153	.83
65+	54.4%	215	
SEX			
MALE	51.2%	165	.35
FEMALE	56.4%	209	
JOINT			
HIP	58.5%	176	.08
KNEE	49.0%	198	

IPC=informed, patient centered

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**Yes, the shared decision making process is the identified healthcare action. It is supported by the stated rationale.
**A conceptual argument is made asserting that the proposed measure assess the quality of treatment decision making for hip or knee replacement. Informed patient-centered decision is defined as the proportion of patients with a passing knowledge score (of the benefits and harms of hip/knee surgery) who stated a preference for surgery (versus non-surgical treatment) for osteoarthritis.

1b. Performance Gap

Comments:

**Yes, the performance data does suggest a gap in shared decision making on the part of informed, patient care surgery. There are suggestions of race and education level disparities as well, but it could be because of the limited sample size. Overall, I support a moderate performance gap ranking, primarily due to the limited sample size.
**This is a new measure. Data on a performance gap are based on parallel constructs in the literature. One study among patients who had discussed hip or knee surgery with their provider (n=141) suggested poor knowledge of harms and benefits. Data provided for disparities using another measure were inconclusive.

1c. PRO-PM

Comments:

**Both clinicians and patients expressed preference in the items utilized in this measure as important to informed and shared decision making.
**Prevalence of osteoarthritis and TJR cited as supporting the measure target as a high priority.

Patient and clinician input in measure development cited as evidence that target population finds measure meaningful. Other evidence supporting measure importance to target population was based on other separate attitudinal studies of physicians and patients.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Decision Quality Instrument (tool), ICD-10 and CPT Codes to identify surgery patients

Specifications:

- The level of analysis for this measure is the clinician group/practice. The measure is specified for the clinician office/clinic setting.
- The numerator is calculated based on patient responses to 6 questions from the Hip or Knee Decision Quality Instruments: five multiple choice knowledge items and one preference item.
    - One point is awarded for each correct knowledge item and then a total knowledge score is calculated and scaled from (0-100%).
    - Respondents who score 60% or higher on knowledge **and** who indicate a clear preference for surgery have a positive decision quality assessment and are counted in the numerator.
    - Those who score less than 60% and/or who are either unclear or prefer nonsurgical options have a negative decision quality assessment, and are not counted in the numerator.
- The denominator includes the number of surveys of patients who have undergone primary knee or hip replacement surgery for osteoarthritis (based on ICD and CPT codes).
- Participants who answer at least 3 of the 5 knowledge items and the preference item will be counted in the denominator (thus, those who do not answer 3 or more of the knowledge items or who do not answer the preference item are excluded from the measure)
- Required codes were submitted. Links to the decision quality instruments for the numerator also were submitted.
- The measure is not risk adjusted nor risk stratified.
- A detailed calculation algorithm is provided.
  Sampling is permitted for this measure, and suggestions for sampling methods are provided. The developer recommends a minimum sample size of 150 responses. Proxy responses are not allowed.

Questions for the Committee :

- *Do you have any specific questions on the specifications, codes, definitions, etc.*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing: Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The sample used for testing included 91 of 382 patients with hip or knee osteoarthritis who were surveyed approximately one year post-surgery or post-consult and who completed a second survey 4-6 weeks later. Respondents were either seen in 1 of 3 clinical sites in the Northeast (one of which used decision aids) or responded to a newspaper advertisement for the research study. The developers reported demographic and clinical characteristics of the test sample.
- Data element reliability: The developer measured test-retest reliability of the knowledge and preference items

from same individuals 4-6 weeks apart. For the knowledge score, the intraclass correlation coefficient (ICC) of the knowledge score at time 1 and time 2 was examined. For the preference item, the developer examined the kappa statistic between the response at time 1 and the response at time 2.

- Score-level reliability: The developer randomly split patients at the same clinical site into groups of 25 or larger and correlated the scores (i.e., this analysis examines how well the score from one sample's reports correlated with another sample's reports for same decision for same provider group). The developer also states that reliability was calculated as variability from site divided by total variability in scores. Correlating results from split samples is an appropriate method of testing the reliability of the measure score. However, the developer does not describe how results from the correlation analysis is used to calculate site-level or total variability, and thus, it is unclear whether this analysis is an appropriate method of testing reliability of the measure score.

Results of reliability testing

- Data element reliability:
  - The ICC from the test-retest analysis of the knowledge score was 0.81 (95% CI 0.71 to 0.87). The ICC reflects the percentage of variance in score results that is due to "true" or real variance between the scores at the 2 time periods. A value of 0.7 is often regarded as a minimum acceptable reliability value.
  - The kappa from the test-retest analysis of the preference item was 0.801. The kappa value represents the proportion of agreement between two raters/abstractors that is not explained by chance alone. A value of 1.0 reflects perfect agreement; a value of 0 reflects agreement that is no better than what would be expected by chance alone. A kappa of 0.801 means that the raters agreed 80.1% of the time over and above what would be expected by chance alone. According to the Landis and Koch classification, this represents "substantial" agreement.
- Measure score reliability: The developer did not report results from the correlation analysis, although they did note that the reliability estimate calculated by dividing the site-level variability by the total variability was 0.853.

Guidance from the Reliability Algorithm

Submitted specifications are precise, unambiguous and complete (Box 1): Yes→ Empirical reliability testing conducted with measure as specified (Box 2): Yes → Reliability testing with computed performance measure score for measured entity/level of analysis (Box 4): Yes →Method Described Appropriate (Box 5): Unclear →TBD if method is appropriate, there is a high level of certainty or confidence that performance measure score is reliable (Box 6):

Questions for the Committee:

- *How were the results of the correlation analysis used to calculate the score-level reliability estimate?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ High ☐ Moderate ☐ Low ☐ Insufficient

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ Yes ☐ Somewhat ☐ No

Question for the Committee:

Do you agree that the specifications consistent with the evidence?

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ Measure score ☐ Data element testing against a gold standard ☐ Both

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method(s):

The validity testing is done both at the individual component level (i.e. knowledge and preferred treatment) and at the measure score level (i.e. informed, patient-centered (IPC) surgery).

- The sample used for testing included 127 patients from an academic medical center in Canada who were referred to an orthopedic surgeon for total joint replacement (TJR) of the hip or knee. These respondents were randomized to receive either a patient decision aid on TJR or usual care. The developers reported demographic and clinical characteristics of the test sample.
- Data element testing:
  - A key feature of a knowledge test is that is can discriminate among those with different levels of knowledge and can detect clinically meaningful differences in knowledge resulting from interventions. As a result, we tested hypotheses that (a) providers would have higher knowledge scores than patients and that (b) patients who had seen a decision aid would have higher knowledge than the control group.
  - The validity of the items used to elicit preferred treatment was evaluated by seeing whether it discriminated patients' ratings of specific goals for pain relief, functional limitations and avoiding surgery. In other words, we examined whether patients who stated a clear preference for surgery rated the importance of relieving pain and improving function higher than those who were unsure or those who stated a preference for nonsurgical treatments. Further, we examined whether those who stated clear preference for surgery rated the importance of avoiding surgery lower than those who were unsure or those who stated a preference for non surgical treatments. These hypotheses were tested using ANOVA with planned comparisons.
  - We tested the predictive validity of the overall IPC surgery measure. We hypothesized that patients who were informed and received treatments that matched their preferred treatment would have higher confidence (using a two sample t-test) and less regret (using a Chi squared test) than those who did not match.
  - We also tested hypotheses that IPC surgery is associated with better health outcomes
- Score-level testing:
  - We tested hypotheses that rates of IPC surgery are higher for patients who report more involvement in decision making process and are seen at a site that has formal decision support processes.

Validity testing results:

- The developer provided data for each level of testing described above and summarized conclusions as follows:
  - Data element testing: The data provide evidence that the measure can discriminate among groups with different levels of knowledge (such as those who have viewed a decision aid or not), and the preference item can discriminate among patients with who place a different amount of importance on salient goals relating to treatment for osteoarthritis.
  - Score-level testing: The IPC surgery measure is significantly higher in practices with formal decision support than in those without formal support (67% vs 51%, p-value < 0.001). Further, the IPC surgery measure demonstrated predictive validity and is associated with higher confidence, less regret and better quality of life.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Other specific question of the validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Respondents who are missing 3 or more knowledge items or the preference item are excluded from the measure.
- The developer examined the frequency of exclusions and also analyzed exclusion patterns across age, sex, education, site, and joint groupings.

Of the 382 respondents who completed surveys, only 8 (2.1%) were excluded (7 did not complete the preference item and 1 did not complete at least 3 of the knowledge items). There were no statistically significant differences between excluded vs. included respondents according to patient characteristics or site. had missing responses for 3 or more knowledge items

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: Risk-adjustment method ☒ None ☐ Statistical model ☐ Stratification

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer compared the measure for practices that had implemented procedures to promote shared decision making and those who did not, including a general population sample. Multivariable logistic regression analyses were used to examine factors associated with rates of informed, patient-centered surgery.
- A randomized controlled trial where the Hip and Knee Decision Quality Instruments were used also provides data on meaningful differences in rates of informed, patient centered surgery for patients who were or were not exposed to patient decision aids.
- Based on the different randomized and non randomized studies, it is possible to see differences from 10%-30% in rates of IPC surgery across sites or groups of patients. From these data we suggest a minimal meaningful difference in scores of 10%.
- *There is considerable evidence that "usual care" results in fairly low rates of IPC surgery, suggesting considerable room for improvement. The evidence is pretty strong that this measure is a valid and reliable assessment of the extent to which patients are well-informed and receive their preferred treatments. The evidence also supports the ability of existing tools (e.g. patient decision aids) to result in a meaningful improvement in the measure.*

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developers considered two ways of handling missing data. The analysis indicate that missing data are uncommon and that choice of how to handle missing data yield similar results for the overall score.

Table: Missing responses and comparison of two approaches for handling missing data for the knowledge items used to generate the measure

Number of questions answered	Frequency (%)	% with Knowledge score 60% or higher (missing as incorrect)	% with Knowledge score 60% or higher (missing with 1/k imputation)
0	1 (0.3%)	0%	0%
1	0 (0%)	n/a	n/a
2	0 (0%)	n/a	n/a

3	2 (0.5%)	0%	0%
4	10 (2.6%)	30%	30%
5	368 (96.5%)	69.5%	69.5%

Guidance from the Validity Algorithm:

Measure specifications consistent with evidence (Box 1): Yes → Relevant potential threats to validity empirically assessed (Box 2): Yes → Empirical validity testing of measure as specified and appropriate (Box 3): Yes →Validity testing of computed performance measure score (Box 6): Yes → Method described and appropriate (Box 7): Yes → Level of certainty or confidence that performance measure scores are a valid indicator of quality (Box 8): High

Preliminary rating for validity: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**I believe the reliability algorithm to be very logical and clear - and it was tested at the correct and consistent level.
**Measure specification is clear except for reporting time period which apparently may vary by reporting unit (clinician: group/practice) to achieve target of 150 patients per center/practice site. Scoring algorithm is clear. Sampling strategy and data collection specifications are vague. Consistency of implementation is therefore unclear.

2a2. Reliability Testing

Comments:

**The sample size in the validity testing seemed small. I cannot speak to the proper method, but the results seem to suggest a moderate level of reliability.
**Reliability was tested on a sample of patients age greater than or equal to 40 with osteoarthritis of hip or knee who had TJR or had discussed surgery with their physician (i.e. did not have surgery) within prior 2 years, who were mailed the survey (n=382) and a subset of these (n=91) who complete a second survey 4 weeks later.

Test-retest reliability on knowledge score was adequate (ICC=0.81), as was preference item (Kappa=0.801).

Practice-level ICC=0.853, although this appeared to be calculated for split-half samples within practices, versus between practice site variation (see p.32).

2b.2 Validity Testing

Comments:

**I do not see that the validity testing was also done at the facility level. But the survey's predictive abilities built some confidence in the measure itself. But the testing was only done at one facility.
**Discriminant validity (comparing patients randomized to receive decision aids versus those not) showed significant differences in knowledge scores favoring decision aid arm. Patients' significant goals for treatment (pain relief, fs) appeared to be related to treatment preference. Higher decision quality was related to decision confidence, lack of decision regret (construct not predictive validity), more shared decision making (measured by?) and greater physical functioning (SF-12 PCS). However, the small sample sizes and the single geographic location severely limit the generalizability of these findings as yet. Since this is a new measure, more empirical work is needed.

2b3.-2b7. Test Related to Potential Threats

Comments:

**I support the exclusion methodology, and although they seem small, I do believe that the statistically significant changes may occur in usual care. Overall, I agree with the high rating for validity.
**2b3. Exclusions did not appear to compromise internal validity of the study. Small sample sizes do limit external validity.

2b4. No risk adjustment was performed.

2b5. Data on meaningful difference appeared to be based on 2 RTCs of decision aids (n=142, n=340). Based on these data, a meaningful difference between sites of 10% in decision quality was suggested. These data must be considered inconclusive until tested in broader, more generalizable settings without intervention trials. No data on between practice variation was provided.

2b7. Missing data appeared to be minimal.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- In item S.21, the developer notes the following:
  - Eligible participants are identified by the clinician, clinical site or third party.
  - The survey has been administered by mail, phone and online for patients to complete at home. The method we have used most often is mail with a postage paid return envelope. A combination of mail and phone reminders are often needed to achieve adequate response rates.
  - A third party vendor may also be used to administer the survey.
- These questions have been extensively cognitively tested to ensure that they are consistently understood and that answers meaningfully describe patient experiences. We have used the questions proposed, and slight variations thereon, in a variety of survey designs: cross-section surveys of adults 40 and older, Medicare beneficiaries known to have had procedures based on claims, and clinical settings in which patients were identified by office staff or via medical records.
- There are no fees for the measure or for the use of the Hip or Knee Decision Quality Instruments used to generate the measure, provided the surveys are used in accordance with the creative commons copyright license.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**The tool will be the most difficult thing to implement on a widespread basis. The data elements are not available in EHRs or any electronic form. Surveys in general are difficult to implement on a national basis. And mailing surveys out will require a lot of work. I am also concerned about getting responses on smaller facilities, with smaller patient populations, as the response rate to surveys can be very small.
**No feasibility assessment provided. No discussion of limitations of mail-out surveys was provided. Use of the measure up to 2 years after surgery may compromise interpretation, patient tracking, etc.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
This measure is used in the Alliance Quality Path recognition program .

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details:

Improvement results:
new measure/no information

Unexpected findings (positive or negative) during implementation:
new measure/none reported

Potential harms:
new measure/none reported

Feedback:
new measure/no information

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4 Usability and Use
Comments:
**The measure is not being publicly reported - but having and supporting informed patient care is important to the quality of care provided across the country. Having informed patients supports the shared decision making that leads higher quality of life for patients - especially on such an important crossroad in care for patients considering surgery.
**Usability is unclear since measure is new.

Criterion 5: Related and Competing Measures

Related or competing measures
None Identified (consider relatedness of 2962 Shared Decision Making Process – also under review)

Harmonization
N/A

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 3/29/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

- ☐ Health outcome: Click here to name the health outcome
- ☒ Patient-reported outcome (PRO): experience with care
  - *PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors*
- ☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

A high quality decision about elective surgery, such as total hip or knee replacement, requires that patients are well-informed and have a clear preference for surgery. The Informed, Patient Centered (IPC) surgery measure presents data on how well centers or hospitals are doing informing patients and tailoring treatments to patients' preferences.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

The measure is a PRO that reflects the quality of the treatment decision making process. The measure reflects multiple care processes and outcomes such as communication, provision of information, shared decision making, and patient engagement.

The use of patient decision aids has been associated with increased decision quality. Further, increased decision quality, and having treatments that match patients' preferences, has been associated with reduced utilization of joint replacement surgery and better health outcomes. [Sepucha et al 2011; Sepucha et al 2013; Stacey et al 2014]

References:

Sepucha K, Stacey D, Clay C, Chang Y, Cosenza C, Dervin G, Dorrwachter J, Feibelmann S, Katz JN, Kearing S, Malchau H, Taljaard M, Tomek I, Tugwell P, Levin C. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. BMC Musculoskelet Disord 2011 Jul 5;12(1):149.

Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.

Stacey D, Légaré F, Col N, Bennett C, Barry M, Eden K, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2014 Jan 28(1).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2958
De.2. **Measure Title:** Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery
Co.1.1. **Measure Steward:** Massachusetts General Hospital
De.3. **Brief Description of Measure:** The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60% or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision.
The target population is adult patients who had a primary hip or knee replacement surgery for treatment of osteoarthritis.
1b.1. **Developer Rationale:** Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

S.4. **Numerator Statement:** The numerator is the number of respondents who have an adequate knowledge score (60% or greater) and a clear preference for surgery.
S.7. **Denominator Statement:** The denominator includes the number of surveys of patients who have undergone primary knee or hip replacement surgery for osteoarthritis. Participants who answer at least 3 of the 5 knowledge items and the preference item will be counted in the denominator.
S.10. **Denominator Exclusions:** Respondents who are missing 3 or more knowledge items do not get a total knowledge score and are not able to be assessed for the measure. Similarly, respondents who do not indicate a preferred treatment do not get counted in the denominator.

De.1. **Measure Type:** PRO
S.23. **Data Source:** Patient Reported Data/Survey
S.26. **Level of Analysis:** Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** The measure is not paired or grouped.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_application_evidence_IPC_Hip_and_Knee_Replacement.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The sample includes patients from three sites and a general population sample from the Boston area. The site that had a formal shared decision making process (SDM site) had a higher rate of informed, patient centered (IPC) surgery than the sites with no formal shared decision making (usual care sites). The association between SDM site and rates of IPC surgery remained significant in multivariate analyses controlling for joint (knee/hip), gender, surgery, and decision making process scores [Sepucha et al 2013].

Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
The DECISIONS study was a national random sample of patients surveyed by telephone up to two years after their decision. They asked earlier versions of four of these knowledge items and found that on the whole, patients had considerable knowledge gaps. For the 141 patients who had discussed hip or knee replacement surgery with their health care provider, the total knowledge score was 32.1% [Fagerlin 2010]. When the researchers combined respondents across different types of elective surgery including back surgery and cataract surgery, race and education were predictors of knowledge (lower education and non White race were associated with lower knowledge).

In summary, data show that patients are not typically well informed about the treatment options for knee and hip replacement surgery, and patients undergo these elective procedures without a clear preference for it. There is considerable room for improvement in elective hip and knee replacement decisions. There is also evidence that

clinical sites that have processes in place to promote share decision making (such as use of patient decision aids) are able to achieve higher rates of IPC surgery than the average or usual care.

Fagerlin A, Sepucha K, Couper M, Levin C, Ubel P, Singer E, Zikmund-Fisher B. Patients' knowledge about 9 common health conditions: Data from a national representative sample. Medical Decision Making Sept/Oct 2010 30: 35S-52S, doi:10.1177/0272989X10378700.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The data come from a sample of patients who were surveyed about one year after surgery or after a visit with an orthopedic surgeon. The covariates we looked at were age (>65, <=65), education (college or more, less than college degree), race/ethnicity (non Hispanic White, other) and gender.

Table: Disparities Data for Knee and Hip Replacement Surgery

VARIABLE	GROUP	IPC	P-value	N
EDUCATION	>=COLLEGE	57.7%	.09	208
	<COLLEGE	48.8%		160
RACE	NON-HISPANIC WHITE	54.5%	.08	352
	OTHER RACES	31.2%		16
AGE	<65	52.9%	.83	153
	65+	54.4%		215
SEX	MALE	51.2%	.35	165
	FEMALE	56.4%		209
JOINT	HIP	58.5%	.08	176
	KNEE	49.0%		198

IPC=informed, patient centered

For the comparison on race/ethnicity, the small number of cases limits the power to detect significant differences.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

Although we did not find significant relationship in this sample between rates of informed, patient-centered surgery and education, there is evidence that less education and non White race are associated with lower knowledge scores (Fagerlin et al, 2010).

Fagerlin A, Sepucha K, Couper M, Levin C, Ubel P, Singer E, Zikmund-Fisher B. Patients' knowledge about 9 common health conditions: Data from a national representative sample. Medical Decision Making Sept/Oct 2010 30: 35S-52S, doi:10.1177/0272989X10378700.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Osteoarthritis (OA) is a leading cause of disability in the U.S. and a growing public health problem. More than one-third of adults 65 and older have OA (1) and the majority report at least some degree of limitation. A significant percentage of patients with OA (40%) report that their overall health is only "fair" or "poor" (2). Studies have also found that adults with OA have higher rates of death from all causes, cardiovascular deaths, and dementia deaths (1.6, 1.7, and 2.0 times higher respectively) compared with the general population (3).

Joint replacement surgery is a common treatment for OA and more than 1,000,000 hip and knee replacements were performed in 2010 in the U.S (4). The decision about whether or not to have joint replacement surgery requires patients and clinicians to make tradeoffs between the chance of symptom relief and potential complications. For example, total hip or knee replacement provides a high likelihood of near complete pain relief (80-90%) but carries a small chance of serious complications (1-5%) and requires considerable time and effort for recovery (5,6). Delaying or waiting for surgery does not decrease the effectiveness, so in order to determine whether or when surgery may be warranted, providers need to understand how bothered patients are by their symptoms and how concerned they are about the prospect of having surgery. The decision to have surgery depends on a complex interplay of having an appropriate clinical condition and patients' informed preferences. Clinical guidelines for treatment of OA emphasize the importance of informing patients and engaging in shared decision making (SDM) to determine the best treatment (5,7).
The Dartmouth Atlas found nearly 10-fold variation in the rates of hip (from 0.6 to 7.5 per 1,000) and knee (from 2.2 to 18.6 per 1,000) replacement procedures for the Medicare population in 2012 (8). The large differences between the high and low rate areas is widely interpreted as evidence that decisions are being driven by providers, not patients, and reflecting highly different physician ideas about how aggressively to use the procedures. Thus, in addition to the large number of procedures involved, this is compelling evidence of a need for greater patient involvement in decision making for these procedures.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
(1) Osteoarthritis. Available at: http://www.cdc.gov/arthritis/basics/osteoarthritis.
(2) Guccione A, Felson D, Anderson J, Anthony J, Zhang Y, Wilson P, et al. The effects of specific medical conditions on the functional limitations of elders in the Framingham Study. Am J Pub Health 1994;84(3):351-358.
(3) Nüesch E, Dieppe P, Reichenbach S, Williams S, Iff S, Jüni P. All cause and disease specific mortality in patients with knee or hip osteoarthritis: population based cohort study. BMJ 2011;342:d1165.
(4) Inpatient Surgery. 2014; Available at: http://www.cdc.gov/nchs/fastats/inpatient-surgery.htm.
(5) Katz JN, Earp BE, Gomoll AH. Surgical management of osteoarthritis. Arthritis Care Res (Hoboken) 2010;62(9):1220-8.
(6) Mantilla CB, Horlocker TT, Schroeder DR, Berry DJ, Brown DI. Frequency of myocardial infarction, pulmonary embolism, deep vein thrombosis, and death following primary hip or knee arthroplasty. Anesthesiology 2002;96(11):40-46.
(7) Jevsevar D. Treatment of osteoarthritis of the knee: evidence-based guideline, 2nd edition. J Am Acad Orthop Surg 2013;21(9):571-576.

(8) Dartmouth Atlas [Accessed on March 10, 2016] http://www.dartmouthatlas.org/tools/downloads.aspx?tab=41

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.**
(Describe how and from whom their input was obtained.)
The development of the items included in the measure was conducted with considerable engagement of patients

and multidisciplinary group of clinicians. Patients and clinicians rated the importance of the items for assessing informed decision making, and the ones included in the measure not only performed well in psychometric analyses but also were rated highly by patient and clinician stakeholders. [Sepucha 2008]

In another study, we surveyed a selected sample (n=279) of primary care clinicians and specialists and asked their opinion about using their patients' knowledge and the extent to which their patients received treatments that match their preferences as quality indicators-the two elements included in the proposed IPC surgery measure. The respondents were considering 14 different medical conditions from cancer screening through elective surgery, and n=50 respondents were specifically considering treatment for knee and hip osteoarthritis. The majority of clinicians were positive (46.8%) or neutral (29.6%) about using their patients´ knowledge as a performance measure and even more were positive (64%) or neutral (22.5%) about using the percentage of patients who received preferred treatments as a performance measure. The responses for hip and knee clinicians were similar to the overall sample.

Although not directly related to hip and knee osteoarthritis, we do have additional evidence that providers generally feel it is important to inform patients and elicit patients' treatment preferences. Providers identified through the American Medical Association master file were surveyed about one of four common decisions: colon cancer screening, herniated disc, menopause, or depression. Overall, 436/737 (59%) of providers responded across the four topics, including 182 primary care physicians (PCPs) and 254 specialists. The respondents were on average 52 years old (SD 9.2), white (73%), male (68%), and had been in practice 21 years (SD 9.5). Almost all providers felt it was very important for their patients to be informed (94% specialists vs. 94.5% PCPs, p=.58). Specialists were more likely to report that their patients were extremely or very well informed compared to PCPs (73% vs. 47%, p<.001). Almost all providers (93%) felt that it was extremely or very important to discuss patients' treatment preferences before a decision is made. Both specialists and PCPs report having such discussions often (98% and 93%, p=0.007). [Sepucha et al 2011]

A recent cross-sectional survey of adults, which was conducted by Public Opinion Strategies, provides additional evidence that patients want to be involved in decision making. Respondents were asked to read a statement about informed decision making (shown below) and rate their favorability toward the concept on a scale from 0 to 100. "Informed medical decision making is an idea in health care that patients should receive information about all of the treatment choices and options available to them for a specific disease, illness, or procedure before they decide, in conjunction with their doctor, on the appropriate treatment choices." With 100 being the most favorable response, the mean rating was 82. Almost 70 percent of respondents rated the statement with a score greater than 80. These data show that when given clearly worded questions about medical decision making, the majority of people want to be involved in an active decision making process.

Citations:

Sepucha K, Levin C, Uzogara E, Barry M, O'Connor A, Mulley A. Developing instruments to measure the quality of decisions: Early results for a set of symptom-driven decisions. Patient Education and Counseling 2008 73:504-510.

Sepucha K, Feibelmann S. What do health care providers think about shared decision making? Presented at the Society for Medical Decision Making annual meeting 2011.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently

within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Musculoskeletal, Musculoskeletal : Joint Surgery, Musculoskeletal : Osteoarthritis

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Patient and Family Engagement, Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.massgeneral.org/decisionsciences/research/DQ_Instrument_List.aspx.

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_IPC_Hip_Knee_Replacement_Measure_ICD10CPTcodes.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
This is a new measure.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator is the number of respondents who have an adequate knowledge score (60% or greater) and a clear preference for surgery.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
There are no set time periods. It would be reasonable for groups to survey patients and report the measure annually, or when they have reached a sufficient volume of responses (minimum recommended number is 150 per center).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator is the number of respondents who have a positive decision quality assessment.

The numerator is calculated based on patient responses to 6 questions from the Hip or Knee Decision Quality Instruments (these items are listed below in S.18 and included as an appendix): five multiple choice knowledge items and one preference item. One point is awarded for each correct knowledge item and then a total knowledge score is calculated and scaled from (0-100%). Respondents who score 60% or higher on knowledge and who

indicate a clear preference for surgery have a positive decision quality assessment and are counted in the numerator. Those who score less than 60% and/or who are either unclear or prefer nonsurgical options have a negative decision quality assessment, and are not counted in the numerator.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator includes the number of surveys of patients who have undergone primary knee or hip replacement surgery for osteoarthritis. Participants who answer at least 3 of the 5 knowledge items and the preference item will be counted in the denominator.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The denominator is all adult patients who had a hip or knee replacement surgery for treatment of osteoarthritis and responded to the Hip or Knee Decision Quality Instrument.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Respondents who are missing 3 or more knowledge items do not get a total knowledge score and are not able to be assessed for the measure. Similarly, respondents who do not indicate a preferred treatment do not get counted in the denominator.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
There is an attached sheet with ICD 10 and CPT codes needed to identify eligible patients to be surveyed for inclusion in the measure.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk stratification used.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

No risk stratification used.

S.16. **Type of score:**
Categorical, e.g., yes/no
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Passing score defines better quality

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The following steps need to be taken to calculate the measure: (1) identify eligible patients (2) administer the Hip or Knee Decision Quality Instrument (3) collect and code responses (4) calculate total knowledge scores and exclude those with 3 or more knowledge items missing (5) calculate the numerator (informed and clear preference for surgery or not) for each individual, excluding those with no knowledge score and/or no preference item and (6) aggregate the measure into a rate over the center or practice.
Responses to five knowledge questions and one preference item from the Hip or Knee Decision Quality Instrument are needed to calculate the Informed, Patient Centered (IPC) surgery measure and are coded and scored as indicated below.

Scoring of Knee Items used to generate the measure
1. Which treatment is most likely to provide relief from knee pain caused by osteoarthritis?
Surgery (Coded- 1)
Non-surgical treatments (coded =0)
Both are about the same (coded= 0)
Multiple responses = 0
Missing response = 0.33

2. After knee replacement surgery, about how many months does it take most people to get back to doing their usual activities?
Less than 2 months (coded= 0)
2 to 6 months (coded = 1)
7 to 12 months (coded= 0)
More than 12 months (coded= 0)
Multiple responses = 0
Missing response = 0.25

3.If 100 people have knee replacement surgery, about how many will have less knee pain after the surgery?
20 (coded= 0)
40 (coded= 0)
60 (coded= 0)
80 (coded = 1)
Multiple response = 0
Missing response = 0.25

4.If 100 people have knee replacement surgery, about how many will have a serious complication within 3 months after surgery?
4 (Coded=1)
10 (coded= 0)
14 (coded= 0)

20 (coded= 0)
Multiple responses = 0
Missing response = 0.25

5. If 100 people have knee replacement surgery, about how many will need to have the same knee replaced again in less than 15 years?
More than half (coded= 0)
About half (coded= 0)
Less than half (coded =1)
Multiple responses = 0
Missing = 0.33

Scoring of Preference Item for Knee:

6. Which treatment did you want to have to treat your knee osteoarthritis?
Surgery (coded=1)
Non-surgical treatments (coded= 0)
Not sure (coded= 0)
Multiple responses (coded=0)

Scoring of Hip Items used to generate the measure:
1. Which treatment is most likely to provide relief from hip pain caused by osteoarthritis?
Surgery (Coded- 1)
Non-surgical treatments (coded =0)
Both are about the same (coded= 0)
Multiple responses = 0
Missing response = 0.33

2. After hip replacement surgery, about how many months does it take most people to get back to doing their usual activities?
Less than 2 months (coded= 0)
2 to 6 months (coded = 1)
7 to 12 months (coded= 0)
More than 12 months (coded= 0)
Multiple responses = 0
Missing response = 0.25

3. If 100 people have hip replacement surgery, about how many will have less hip pain after the surgery?
30 (coded= 0)
50 (coded= 0)
70 (coded= 0)
90 (coded = 1)
Multiple response = 0
Missing response = 0.25

4. If 100 people have hip replacement surgery, about how many will have a serious complication within 3 months after surgery?
4 (Coded=1)
10 (coded= 0)
14 (coded= 0)
20 (coded= 0)
Multiple responses = 0

Missing response = 0.25

5. If 100 people have hip replacement surgery, about how many will need to have the same hip replaced again in less than 20 years?
More than half (coded= 0)
About half (coded= 0)
Less than half (coded =1)
Multiple responses = 0
Missing = 0.33

Scoring of Preference Item for Hip:
6. Which treatment did you want to have to treat your hip osteoarthritis?
Surgery (coded=1)
Non-surgical treatments (coded= 0)
Not sure (coded= 0)
Multiple responses (coded=0)

Knowledge: The responses are coded as indicated above. A total knowledge score is calculated by summing the five items, dividing by 5 and converting to percentage to get scores 0-100%. Missing answers are imputed with 1/k where k is the number of possible responses (essentially equivalent to guessing). Multiple responses (e.g. on paper survey) are considered incorrect and coded as 0. A total knowledge score is calculated for all surveys that have three or more knowledge items completed.

Preference item: Respondents who mark surgery are considered to indicate a clear preference for surgery. Respondents that mark either non surgical treatments or not sure, are not considered to have a clear preference for surgery. Missing responses are not counted. Multiple responses (e.g. on a paper survey) are considered "not sure" and coded as 0.

A positive assessment "yes" for decision quality requires a knowledge score of 60% or higher and a clear preference for surgery. Otherwise, decision quality is "no."

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Patients of a particular surgeon or at a particular clinical site (which could be a group of providers or a hospital or other surgical site) who had a primary knee or hip replacement surgery are identified from medical records, claims or in some other way. Patients can be sampled sequentially, or a pool of such patients who had the procedure in a particular time period (e.g. in the last 12 months) can be created and sampled at a rate that produces the desired number of potential respondents.

The Decision Quality Instruments from which the measure is calculated can be used in a population-based sample, such as a sample of a population in a geographic area. Eligible respondents could be identified from claims (such as Medicare claims files) or based on patient self- reports of having had the procedures within some time frame.

The Decision Quality Instruments have also been used with patients shortly after a consult with an orthopedic

surgeon to discuss joint replacement surgery. However, there is often not consistent or detailed enough coding of visits to reliably identify patients after the visit but before having one of these procedures. As a result, at this time, the measure is proposed for use with patients who have had surgical treatment.

For knee and hip replacement surgery, rates of informed, patient-centered surgery varied from 37.9% to 59.5% across sites. A general population sample of patients who had knee and hip replacement surgery had rates of informed, patient-centered surgery of 18.8%. A sample size about 150 would be needed to detect differences in proportions of 15% for the measure (e.g. from 25% to 40%) with 80% power. This size difference is what we have observed between sites that do and do not make an effort to do shared decision making.
Proxy respondents are not permitted. The patients who receive the procedure should answer the survey questions. The survey is available in English and Spanish.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Eligible participants are identified by the clinician, clinical site or third party.
The survey has been administered by mail, phone and online for patients to complete at home. The method we have used most often is mail with a postage paid return envelope. A combination of mail and phone reminders are often needed to achieve adequate response rates.
A third party vendor may also be used to administer the survey.
We recommend that data not be accepted if response rates are lower than 50%. Calculate response rate as all those responding divided by all those invited to answer the survey questions (American Association for Public Opinion Research (AAPOR) response rate 4).

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
If one or two knowledge items are missing, assign a code of 0. It is assumed that if one or two items are skipped, then the respondent does not know the correct answer. If more than two knowledge items are missing then delete the case.
If the preference item is missing then delete the case.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Patient Reported Data/Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The measure is derived from responses to the Hip and Knee Decision Quality Instruments. These patient reported surveys have been administered by mail, phone, and online for patients.
The method we have used most often is mail with a postage paid return envelope. A combination of mail, email, and phone reminders are often needed to achieve adequate response rates.
A third party vendor may also be used to administer the survey.
We have used these questions in English and Spanish.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)* Ambulatory Care : Clinician Office/Clinic If other:
S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)* Not applicable.
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** nqf_testing_attachment_IPC_Hip_and_Knee_Replacement.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Informed, Patient Centered (IPC) Knee and Hip Replacement Surgery

Date of Submission: 3/29/3016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores

are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Patient reports Knee and Hip Osteoarthritis Decision Quality Instrument survey data	☒ other:Patient reports Knee and Hip Osteoarthritis Decision Quality Instruments survey data

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Sample 1: A sample of 382 patients with hip and knee osteoarthritis were surveyed about one year after surgery or one year after discussing surgery with a surgeon. The respondents came from 3 different clinical sites in the Northeast, one of which was using decision aids and encouraging shared decision making for joint replacement surgery, a fourth group was general population sample who responded to

a newspaper ad for the research study. A subset of respondents was sent the same survey 4-6 weeks later to examine retest reliability.

Sample 2: A sample of 127 patients who were part of a randomized controlled trial of knee and hip osteoarthritis patient decision aids were used to examine discriminant validity of the knowledge component of the measure.

1.3. What are the dates of the data used in testing? 2009-2010

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Sample 1: Participants were selected from three academic medical centers in the Northeast and from the community. The community sample responded to an advertisement in a local newspaper.

Sample 2: Participants were selected from an academic medical center in Canada that was running a randomized controlled trial of hip and knee osteoarthritis decision aids.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Sample 1: The full sample included n=382 (79% response rate to mailed survey) and a subset n=91 (83% response rate) completed the retest survey about 4 weeks after the initial survey. Respondents were aged 40 years and older with a diagnosis of hip or knee osteoarthritis who either had total joint replacement or had discussed surgery with their physician (and chosen not to have TJR), within the past two years. Individuals with rheumatoid arthritis, psoriatic arthritis, osteonecrosis, partial knee replacement, revision surgery, or bilateral knee surgery were excluded.

Sample 2: The full sample included 127 respondents (92% response rate to the phone survey). Adult patients with osteoarthritis of the hip or knee who met the guidelines for referral to an orthopaedic surgeon for total joint replacement (TJR) and had access to a TV with a VCR or DVD player were recruited for participation. Patients with inflammatory arthritis; a previous total joint replacement; or who were

deaf, blind, cognitively impaired, or had a language barrier were excluded. After signing a consent form, patients were randomized to receive either a patient decision aid on TJR or usual care. Both groups were instructed to review the information at home and complete the decision quality survey items. Approximately one week after recruitment, a research assistant telephoned participants to record the answers. The research assistant made an average of four calls to participants to complete the survey.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Sample 1 was used for reliability. Samples 1 and 2 were used for validity.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Table 1: Demographic characteristics of patient respondents for each study.

	Sample 1	Sample 2	
Characteristic	All patients N = 382	Hip/Knee Control N=66	Hip/Knee PtDA N=61
Gender: Male n (%)	169 (44)	27 (40.9)	25 (40.9)
Age mean (SD)	62.7 (9.6)	66.1 (9.49)	64.3(10.16)
Race/ethnicity n (%)			
White	359 (95.5)	Not asked	Not asked
Education n (%)			

≥ College graduate	209 (56)	40 (60.6)	39 (63.9)
Some college	94 (25.2)	Not asked	Not asked
High school or less	68 (18.1)	26 (39.4)	22 (36.1)
Missing	9 (2.4)	0	0
Income n (%)			
<$30,000	78 (20.5)	5 (7.6)*	7 (11.5)*
$30,000-60,000	70 (18.3)	21 (31.8)**	18 (29.5)**
$60,000-100,000	89 (23.3)	13 (19.7)	21 (34.4)
Over $100,000	93 (24.3)	22 (33.3)	12 (19.7)
Missing	52 (13.6)	5 (7.6)	3 (4.9)
Married/Committed relationship n (%)	255 (67.8)	42 (63.6)	38 (62.3)
Months since decision median (IQR)	11 (7, 15)	Considering decision	Considering decision
Had (or preferred) Surgery n (%)	235 (61) Had surgery	49 (74.2) Preferred surgery	39 (63.9) Preferred surgery

Joint (knee vs. hip): **Knee n (%)**	201 (53)	61 (94)	59 (97)
WOMAC Pain Score **mean (SD)**	5.6 (4.6)	10.7 (4.2)	11.2 (4.0)

PtDA=decision aid group; SD=standard deviation; N/A=not asked; FT=fulltime; IQR: interquartile range; * measured < $20,000; ** measured from $20,000; WOMAC=Western Ontario McMasters University Arthiritis Index is a measure of disease specific pain

---

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

1. At the item level, we measured test-retest reliability of the knowledge and preference items from same individuals 4-6 weeks apart. For the knowledge score we examined the intraclass correlation coefficient (ICC) of the knowledge score at time 1 and time 2. The ICC compares the variability of different ratings of the same subject to the total variation across all ratings and all subjects. For the preference item, we examined the kappa between the response at time 1 and response at time 2. The kappa statistic measures agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, since κ takes into account the agreement occurring by chance.
2. At the practice level, we randomly split patients at the same clinical site into groups of 25 or larger and correlated the scores; i.e. how well score from one sample's reports correlated with another sample's reports for same decision for same provider group. The reliability was calculated as variability from site divided by total variability in scores.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

1. The test-retest reliability of the knowledge score was examined in sample 1 and found to be ICC=0.81 (95% CI 0.71 to 0.87). The test-retest reliability of the item assessing preferred treatment was (Kappa = 0.801).
2. At the practice level, the reliability was 0.853 (variability from site divided by total variability).

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test-retest reliability for the knowledge and preference items used to generate the measure is high. The reliability of the measure at the clinical practice level is also strong.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The analyses replicate those published in Sepucha et al 2011 [1] and Sepucha et al 2013 [2] using the definition of the informed, patient centered hip and knee replacement surgery measure proposed here. The validity testing is done both at the individual component level (i.e. knowledge and preferred treatment) and at the measure level (i.e. informed, patient-centered (IPC) surgery).

(1) A key feature of a knowledge test is that is can discriminate among those with different levels of knowledge and can detect clinically meaningful differences in knowledge resulting from interventions. As a result, we tested hypotheses that (a) providers would have higher knowledge scores than patients and that (b) patients who had seen a decision aid would have higher knowledge than the control group. Tested using two sample t-tests.

(2) The validity of the item used to elicit preferred treatment was evaluated by seeing whether it discriminated patients' ratings of specific goals for pain relief, functional limitations and avoiding surgery. In other words, we examined whether patients who stated a clear preference for surgery rated the importance of relieving pain and improving function higher than those who were unsure or those who stated a preference for nonsurgical treatments. Further, we examined whether those who stated clear preference for surgery rated the importance of avoiding surgery lower than those who were unsure or those who stated a preference for non surgical treatments. These hypotheses were tested using ANOVA with planned comparisons.

(3) We tested the predictive validity of the overall IPC surgery measure. We hypothesized that patients who were informed and received treatments that matched their preferred treatment would have higher confidence (using a two sample t-test) and less regret (using a Chi squared test) than those who did not match.

(4) We tested hypotheses that rates of IPC surgery are higher for patients who report more involvement in decision making process and are seen at a site that has formal decision support processes. We also tested hypotheses that IPC surgery is associated with better health outcomes. We first examined the following factors: age (<60 years vs_60 years), education (college or more vs other), sex,treatment (surgery vs nonsurgery), joint (hip vs knee), site, quality of life (SF-12 physical component score), and decision process score in univariate analyses using chi-square or t-tests, as appropriate. Then we developed a multivariable logistic regression model with high IPC surgery (yes/no) as the dependent variable and included all variables that were p<0.1 on univariate analyses as independent variables.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

1. We examined discriminant validity of the knowledge assessment by comparing scores of those who should have higher knowledge (e.g. scores of patients who had used a decision aid versus those who did not.) The mean knowledge scores discriminated between patients in decision aid group 67% (SD 21.2) compared to 51% (SD 24.9) in the usual care group (p<0.001). [Sepucha et al 2010]

2. To establish validity, we examined the extent to which patients' stated preference varied appropriately with specific goals. The table below provides evidence of the relationships in the predicted directions, supporting the validity of the single item as reflecting patients' preferred treatment.

On a scale of 1 to 10 where 1 is not at all important and 10 is extremely important,	Prefer surgery N=218	Unsure N=26	Prefer non surgical treatments N=126	P (ANOVA)
How important is it to relieve your knee pain?	9.50 (SD1.19)	8.92 (SD 1.47)	8.43 (2.42)	F=10.87 P<0.001
How important is it not to be limited in what you can do because of your knee pain?	9.74 (SD 0.79)	9.38 (SD 1.33)	8.82 (1.92)	F=12.37 P<0.001
How important is it to you to avoid having surgery?	3.21 (SD 3.18)	5.50 (SD 2.92)	7.96 (SD 2.33)	F=71.65 P<0.001

3. Respondents had met criteria for decision quality were more confident in their decision (9.09/10 vs. 7.78/10, p<0.001) and were significantly more likely to say they would do the same thing again (59.9%% vs. 26.4%%, p<0.001).

4. Replicating the multivariable logistic regression analyses from Sepucha 2013 [2] with the IPC surgery measure as proposed here, found the same results. None of the patient factors (age, sex, education) were significantly associated with IPC surgery. Controlling for treatment, IPC surgery was associated with more shared decision making and with the site that used decision aids. Further IPC surgery was significantly associated with higher quality of life as measured by the SF-12 Physical Component Score. Table below contains the results of these analyses.

Table: Results multivariate logistic regression with IPC surgery as dependent variable.

Variable	Odds Ratio	95%CI	P
Had Surgery	2.462	1.45, 4.17	.001
Site (newspaper)	referent		.016
Site 1	.896	.38, 2.10	.800

Site 2 (decision aid site)	2.275	1.22, 4.25	.010
Site 3	1.500	.69, 3.25	.305
Quality of life (SF-12 Physical component score)*	1.037	1.01, 1.06	.003
Shared decision making score*	1.012	1.00, 1.02	.015
College_grad	1.110	.67, 1.84	.686
Constant	.045		.000

*Odds ratio for a 10-point increase in scores.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The data provide evidence that the measure can discriminate among groups with different levels of knowledge (such as those who have viewed a decision aid or not), and the preference item can discriminate among patients with who place a different amount of importance on salient goals relating to treatment for osteoarthritis.

The IPC surgery measure is significantly higher in practices with formal decision support than in those without formal support. Further, the IPC surgery measure demonstrated predictive validity and is associated with higher confidence, less regret and better quality of life.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

The IPC hip and knee replacement measure excludes surveys that have 3 or more knowledge responses missing or the preference item missing. To evaluate how exclusions might affect validity we examined the frequency of *included* and *excluded* responses across patient characteristics including age, sex, education, and joint (hip or knee). To perform this analysis we created frequency distribution tables then performed a chi-square goodness of fit test. The chi-square tests the null hypothesis that there are no significant differences in the amount of included or excluded surveys between groups. If the test is significant to a p-value of 0.05 or less then we reject the null hypothesis and conclude there are significant differences between groups.

To evaluate the effect of exclusions across organizations, we examined the frequency of included and excluded responses for each site and tested for difference using a chi-square test.

We also calculated "expected" cell frequencies. The expected cell frequency represent the expected frequency of responses should the null hypothesis be true. This allows us to evaluate the departure from the expected number of excluded responses under the null hypothesis.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

We found very little missing data and as a result, there were very few exclusions. In sample 1, 2.1% or 8/382 respondents were excluded for not completing enough items.[1] Of those 8 exclusions, 7/8 did not complete the preference item and 1/8 did not complete at least 3 of the knowledge items.

Table: Included and excluded responses by characteristic with expected frequencies.

Variable (chi-square p-value)	Included (Expected) Column %	Excluded (Expected) Column %
Age (p=0.41)		
Age >65	153 (151.5) 41.6%	1 (2.5) 16.7%
Age <65	215 (216.5) 58.4%	5 (3.5) 83.3%
Joint (p=0.49)		
Hip	176 (177.2) 47.1%	5 (3.8) 62.5%
Knee	198 (196.8) 52.9%	3 (4.2) 37.5%
Sex (p=0.74)		
Male	165 (165.5) 44.1%	4 (3.5) 50%
Female	209 (208.8) 55.9%	4 (4.5) 50%
Education (p=0.17)		
College or more	208 (206.2) 56.5%	1 (2.8) 20%
Less than college	160 (161.8)	4 (2.2)

	43.5%	80%
Site (p=0.82)		
Site 1	50 (49.9) 13.4%	1 (1.1) 12.5%
Site 2	173 (174.3) 46.3%	5 (3.7) 62.5%
Site 3	66 (65.6) 17.6%	1 (1.4) 12.5%
Site 4	85 (84.2) 22.7%	1 (1.8) 12.5%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Overall, we found very few exclusions. We did not find any significant differences by site or by patient characteristics; however, with this sample size there was limited power to detect significant differences. Even if there were some statistically significant differences, the magnitude is likely to be very small so that the effect of those differences on results would be minimal and not likely sufficient to bias results.

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

We do not recommend risk adjustment for this measure. Any patient who has one of these elective surgeries, should be able to answer the knowledge questions correctly (to meet the standards of informed consent) and should have a clear preference for the procedure.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
No risk adjustment.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We compared the measure for practices that had implemented procedures to promote shared decision making and those who did not, including a general population sample. Multivariable logistic regression analyses were used to examine factors associated with rates of informed, patient-centered surgery.

A randomized controlled trial where the Hip and Knee Decision Quality Instruments were used also provides data on meaningful differences in rates of informed, patient centered surgery for patients who were or were not exposed to patient decision aids.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark,*

different from expected; how was meaningful difference defined)

There was considerable variation in rates of IPC surgery across sites, (31.8%, 50.0%, 56.0%, 64.7%) and in all cases, there was considerable room for improvement in rates. Compared to the general population referent group, the site that used patient decision aids achieved significantly higher rates of IPC OR 2.275 (95% CI 1.22, 4.25) [2].

Two randomized controlled trials provide additional evidence for the potential magnitude of impact of decision aids on rates of IPC surgery. In the first, a randomized controlled trial with 142 patients found higher rates of IPCS surgery in the intervention (patient decision aid) compared to control (pamphlet) group (56.4% intervention versus 25.0% control; $p < 0.001$). [3] In the second, a randomized controlled trial evaluating the same decision aids with 340 patients, rates of IPC surgery were also higher in the intervention (56.1%) compared to the control group (44.5%), relative risk (RR) 1.25; 95% CI 1.00-1.56, P = 0.050.[4]

Based on the different randomized and non randomized studies, it is possible to see differences from 10%-30% in rates of IPC surgery across sites or groups of patients. From these data we suggest a minimal meaningful difference in scores of 10%.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

There is considerable evidence that "usual care" results in fairly low rates of IPC surgery, suggesting considerable room for improvement. The evidence is pretty strong that this measure is a valid and reliable assessment of the extent to which patients are well-informed and receive their preferred treatments. The evidence also supports the ability of existing tools (e.g. patient decision aids) to result in a meaningful improvement in the measure.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

_____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

We considered different approaches for handling missing data for the knowledge items. The first approach is to consider a missing answer as incorrect (with those responses coded as 0). The second approach is to impute the score of 1/k where k is the number of potential response options (essentially providing the points equivalent to guessing from the available multiple choice responses). We calculated the frequency of missing responses for each item in the knowledge assessment and then conducted sensitivity analyses to examine the impact on total knowledge scores.

As described earlier in section 2b3, 7/382 (1.8%) of respondents did not complete the preferred treatment item. We exclude respondents who do not complete that item and presented the results of those analyses in the earlier section.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; <u>if no empirical sensitivity analysis</u>, identify the approaches for handling missing data that were considered and pros and cons of each*)

The Table below shows the overall frequency of missing data for individual knowledge items. Twelve participants (3.1%) had 1 or 2 items missing and one respondent did not complete any items (0.3%). The knowledge scores are considerably lower for respondents with missing data; however the samples are very small.

Table: Missing responses and comparison of two approaches for handling missing data for the knowledge items used to generate the measure

Number of questions answered	Frequency (%)	% with Knowledge score 60% or higher (missing as incorrect)	% with Knowledge score 60% or higher (missing with 1/k imputation)
0	1 (0.3%)	0%	0%
1	0 (0%)	n/a	n/a
2	0 (0%)	n/a	n/a
3	2 (0.5%)	0%	0%
4	10 (2.6%)	30%	30%
5	368 (96.5%)	69.5%	69.5%

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing*

data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data)

Generally, missing data are low. Given the threshold for the indicator variable (correctly answering three or more items), the approach to missing data (either imputing 1/k or considering it incorrect) does not impact the % of respondents who meet that threshold. As a result, missing data and the approach to treating missing data have a negligible impact on the rates of IPC surgery.

Citations:

1. Sepucha K, Stacey D, Clay C, Chang Y, Cosenza C, Dervin G, Dorrwachter J, Feibelmann S, Katz JN, Kearing S, Malchau H, Taljaard M, Tomek I, Tugwell P, Levin C. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. BMC Musculoskelet Disord 2011 Jul 5;12(1):149.
2. Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.
3. Stacey D(1), Hawker G, Dervin G, Tugwell P, Boland L, Pomey MP, O'Connor AM, Taljaard M. Decision aid for patients considering total knee arthroplasty with preference report for surgeons: a pilot randomized controlled trial. BMC Musculoskelet Disord. 2014 Feb 24;15:54. doi: 10.1186/1471-2474-15-54.
4. Stacey D(1), Taljaard M(2), Dervin G(3), Tugwell P(4), O'Connor AM(5), Pomey MP(6), Boland L(7), Beach S(8), Meltzer D(9), Hawker G(10). Impact of patient decision aids on appropriate and timely access to hip or knee arthroplasty for osteoarthritis: a randomized controlled trial. Osteoarthritis Cartilage. 2016 Jan;24(1):99-107. doi: 10.1016/j.joca.2015.07.024. Epub 2015 Aug 4.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Other

If other: patient reported outcome

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

No data elements are in defined fields in electronic sources

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.
The data are from patient self report. The surveys can be administered online to support electronic capture but we have found highest response rates to mailed surveys.

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
These questions have been extensively cognitively tested to ensure that they are consistently understood and that answers meaningfully describe patient experiences. We have used the questions proposed, and slight variations thereon, in a variety of survey designs: cross-section surveys of adults 40 and older, Medicare beneficiaries known to have had procedures based on claims, and clinical settings in which patients were identified by office staff or via medical records. The following observations have informed this proposal.
1. While we have included an "I am not sure" response with the knowledge items, particularly when used in the clinic at the time of initial decision making, when we have removed that option, the knowledge scores are higher as many patients do have a sense of the correct answer and will indicate it.
2. We can identify patients making decisions by asking them whether or not they had discussed an intervention, test or treatment. However, for cross-sections of adults or patients, the rates of any particular decision being made are too low to produce reliable data without very large samples.
3. We have surveyed patients in clinical settings before they had treatment. That is certainly the preferred way to measure informed, patient-centered surgery at a clinical site. However, it requires considerable integration into the clinic workflow and significant resources to get adequate response rates. It is easier to accomplish at sites that routinely assess patient-reported outcomes for all surgical patients (as the Decision Quality Instrument items can be included as part of the pre-operative assessment). It is also easier at sites that routinely use patient decision aids for their hip and knee osteoarthritis patients. In order to get comparable results across clinicians or clinical sites, we recommend sampling those patients who actually had the target intervention. In that way, patients can be reliably identified.
4.The hip and knee results are similar within sites, and as a result, we feel that it is reasonable to combine these two decisions in this measure.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees for the measure or for the use of the Hip or Knee Decision Quality Instruments used to generate the measure, provided the surveys are used in accordance with the creative commons copyright license.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Professional Certification or Recognition Program The Alliance Quality Path Program http://www.the-alliance.org/uploadedFiles/Providers/QualityPath_knee_and_hip_replacement_RFP.pdf Quality Improvement (Internal to the specific organization) Shared Decision Making Program http://www.massgeneral.org/decisionsciences/

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Quality Path Program sponsored by the Alliance specifies measurement of shared decision making as part of their criteria for recognition. The Alliance is a cooperative of employers that includes more than 240 members who provide self-funded health benefits to more than 100,000 individuals. The network lets members choose from more than 80 hospitals, 13,500 total professional service providers, and 3,400 medical clinic sites in Wisconsin, Illinois, and Iowa. The purpose of the Quality Path program is to recognize providers and hospitals who are delivering high quality surgical care. The relevant section from the program detailing use of the measure is excerpted below and the entire program details can be found at the website link listed above.
16 Decision Quality Assessment (p 18)
Supporting Documentation:
• Provide a description of the process for assessing the quality of shared decision making. This process needs to use the decision quality assessment tool available at:
o Knee: http://www.massgeneral.org/decisionsciences/assets/pdfs/OAKnee_DQI_SV.pdf
o Hip: http://www.massgeneral.org/decisionsciences/assets/pdfs/OAHip_DQI_SV.pdf
• Ideally, for each procedure, provide percentages, numerators, and denominators of patients participating in an assessment of shared decision making broken out by physician, practice, and by facility. Denominator is all patients receiving elective knee replacement or elective hip replacement. We are looking for reporting capability and evidence of process implementation. If the process has not been in place long enough to produce these numbers, this requirement may be waived until the six-month maintenance of designation process.

The Shared Decision Making Program sponsored in part by Partners Healthcare and Massachusetts Physician´s Organization has a program in place to survey patients with hip and knee osteoarthritis who come to see an orthopedic surgeon. The survey is included in the patient reported outcomes registry that new patients complete and responses are shared with surgeons.

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
It is a new measure and has not been available to be included in public reporting or other accountability applications.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for

implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

It is a new measure.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We have not encountered unintended negative consequences to use of the measure. We have, on occasion, sent patients the correct answers to the knowledge items as we did receive questions about that from respondents who did not learn this information from their health care providers.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** IPC_Hip_and_Knee_Measure_Version.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Massachusetts General Hospital
Co.2 **Point of Contact:** Karen, Sepucha, ksepucha@mgh.harvard.edu, 617-724-3350-
Co.3 **Measure Developer if different from Measure Steward:** Massachusetts General Hospital
Co.4 **Point of Contact:** Karen, Sepucha, ksepucha@mgh.harvard.edu, 617-724-3350-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This was not the product of a formal work group. It emerged from an ongoing effort to develop measures of decision quality since 2007 led by Dr. Karen Sepucha of Massachusetts General Hospital and colleagues at the Informed Medical Decisions Foundation. The following researchers played a significant role at one or more points in the development process.

Karen Sepucha, at Massachusetts General Hospital, has been leading the development efforts team in this area.
Carol Cosenza at the Center for Survey Research at UMass Boston has worked on cognitive testing of these questions in various forms.
Floyd J Fowler, Jr and Carrie Levin at the Informed Medical Decisions Foundation have played the roles in development and testing of the surveys.

Jeffrey Katz, MD Brigham and Women's Hospital provided clinical input into the evidence for the knowledge items.

Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** 2016 Ad.3 **Month and Year of most recent revision:** 01, 2016 Ad.4 **What is your frequency for review/update of this measure?** As needed, at least every two years. Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017
Ad.6 **Copyright statement:** Copyright holder of the Hip and Knee Decision Quality Instruments used to generate the measure is Massachusetts General Hospital (MGH). MGH makes the survey available for use free of charge under the creative commons license agreement, with the provision it is not modified or sold. Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2962
Measure Title: Shared Decision Making Process
Measure Steward: Informed Medical Decisions Foundation, a division of Healthwise
Brief Description of Measure: This measure assesses the extent to which health care providers actually involve patients in a decision-making process when there is more than one reasonable option. This proposal is to focus on patients who have undergone any one of 7 common, important surgical procedures: total replacement of the knee or hip, lower back surgery for spinal stenosis of herniated disc, radical prostatectomy for prostate cancer, mastectomy for early stage breast cancer or percutaneous coronary intervention (PCI) for stable angina. Patients answer four questions (scored 0 to 4) about their interactions with providers about the decision to have the procedure, and the measure of the extent to which a provider or provider group is practicing shared decision making for a particular procedure is the average score from their responding patients who had the procedure.
Developer Rationale: We have collected a great deal of data from cross-section surveys of patients who have made decisions and from patients drawn from clinical sites documenting that for many decisions, patients routinely do not perceive that they discuss the cons of proposed interventions, are not told about alternatives and are not asked for their input into the decisions. Consistently, their levels of knowledge of information relevant to the decisions they are making are low. We then have evidence that when clinicians commit to shared decision making, by routinely providing decision aids for example, the scores of patients with respect to knowledge and the decision making process are higher. We believe the use of the Shared Decision Making Process Score and appropriate measures of patient knowledge can be catalysts to routinely informing and involving patients in important medical decisions, which in turn will increase the likelihood that patients will get the care they want and that is consistent with their goals and concerns (Stacey et al 2014).

Numerator Statement: Patient answers to four questions about whether not 4 essential elements of shared decision making (laying out options, discussing the reasons to have the intervention and not to have the intervention, and asking for patient input) were part of the interactions with providers when the decision was made to have the procedure.
Denominator Statement: All responding patients who have undergone one of the following 7 surgical procedures: back surgery for a herniated disc; back surgery for spinal stenosis; knee replacement for osteoarthritis of the knee; hip replacement for osteoarthritis of the hip; radical prostatectomy for prostate cancer; percutaneous coronary intervention (PCI) for stable angina, and mastectomy for early stage breast cancer.
Denominator Exclusions: : For back, hip, knee, and prostate surgery patients, there are no exclusions, so

long as the surgery is for the designated condition.

PCI patients who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).

For patients who have mastectomy, patients who had had a prior lumpectomy for breast cancer in the same breast and patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies) are excluded.

Measure Type: PRO
Data Source: Patient Reported Data/Survey
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence: Shared Decision Making Process is a Patient Reported Outcome Performance Measure (PRO-PM), as such the developer is required to state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

- The developer states: *When faced with a medical problem for which there is more than one reasonable approach to treatment or management, shared decision making means providers should outline for patients that there is a choice to be made, discuss the pros and cons of the options and make sure that patients have input into the final decision. The result will be decisions that align better with patient goals, concerns and preferences.*
- In addition, the rationale supporting the measure is noted as: *When physicians provide balanced information to patients (often in the form of decision aids) and have a discussion about the options and about what patients want, patients answers these questions in a way that reflects a shared decision making process.*
- *As PRO-PM the developer was asked to provide evidence that the target population values the measured PRO and finds it meaningful. They cited three studies, among them:*
    - *Every time we have surveyed patients about their support for the use of decision aids and how they want to be involved in medical decision making, the results have been overwhelming support for both. For example, data from 2800 patients who were given decision aids when faced with medical decisions found that 83% said that it was "extremely" or "very" important that patients receive decision aids when making decisions like the ones they made. (Wexler,et al., 2015).*

Question for the Committee:

- *Do you agree that there is at least one clinical action is identified and supported by the rationale?*

Guidance from the Evidence Algorithm
PRO-based measure (Box 1) → Relationship between the outcome and at least one healthcare action is identified and supported by the rationale (Box 2) → PASS

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer indicates the best data on the opportunity and need for improvement come from national surveys of patients who either were known to have had procedures, based on Medical claims, or said they had made decisions with their doctors. They provide an attachment that describes mean performance scores on the measure from national surveys, as well as clinical practices.
- The developer provided data comparing clinical sites that have made a commitment to do shared decision making with the shared decision making process scores in usual care, derived both from cross-sectional surveys of patients who have made the decisions or clinical sites that were making no special effort to implement shared decision making.
- The data in the five tables in the attachment consistently show that clinical sites that made a special effort to implement shared decision making have significantly significant higher shared decision making process scores from their patients than patients in "usual care".

The following table summarizes means and SDs for scores from national samples (please reference the linked attachment for all 5 tables):

Table 1: Mean Shared Decision Making Process Scores for 7 Common Surgical Procedures

Procedure	Mean Shared Decision Making Process Score	Standard Deviation	Data Source
Prostatectomy for Prostate Cancer	2.7	1.0	Survey of Medicare Patients who had surgery
Mastectomy for Breast Cancer	1.9	1.3	Survey of Medicare patients who had had surgery
PCI for coronary artery	1.2	1.0	Survey of Medicare patients who had

disease			had surgery
Hip replacement for osteoarthritis of the hip	2.5	1.2	National survey of adults 40 or older from Knowledge Networks panel (TRENDS)
Knee replacement for osteoarthritis of the knee	2.8	1.1	National survey of adults 40 or older from Knowledge Networks panel (TRENDS)
Surgery for lower back pain (disco or stenosis)	3.2	1.0	National survey of adults 40 or older from Knowledge Networks panel (TRENDS)

Note: 4 is highest score attainable.

Disparities

- The developer indicates the availability of data comparing reported shared decision making process scores by patient age, gender, education and race. Although there are some examples of significant relationships in the data, they do not go in consistent directions. The takeaway from the data is there is no evidence that the processes of decision making with providers are consistently related to any of those patient demographic characteristics. These results are presented in detail in NQF table attachment.
- The developers also note that although there is consistent evidence that patient levels of formal education are related to measures of patient knowledge (e.g Fagerlin et al, 2010) there is not sufficient evidence that racial or education groups consistently differ in the their reported interactions with providers about surgical decisions.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:
**If measuring a health outcome or PRO: is the relationship between the measured outcome/PRO and at least one healthcare action (structure, process, intervention, or service) identified AND supported by the stated rationale?

The Shared Decision Making Process Score is a PRO to assess extent of clinician interaction with patients regarding shared decision making for patient decision to have 1 of 7 surgical procedures. The rationale identifies that the score generated from a 4 item questionnaire relates specifically to processes by which a clinician discusses with the patient: options available for treatment of their condition, pros and cons of the intervention (surgery) and general "patient input".

**The proposed measure provides information about a healthcare processes from the patient perspective.

Involving patients in shared decision making is presumed to increase the likelihood that patients receive care that is responsive to their goals and concerns. This is an important patient outcome in its own right. Is there evidence that SDM improves patient satisfaction or other patient outcomes?

1b. Performance Gap
Comments:
**Developer provides data from Healthwise surveys and Dartmouth Atlas showing wide variability in prevalence of the 7 targeted procedures, which they note is an indication of lack of SDM and a provider driven system. They also cite survey data indicating gaps in both patient knowledge and actions associated with shared decision making (discussion of options, sharing decision aids) as rationale for a national measure.

Questions for developer: The proposed process measure does not include a question to assess presence or use of decision aids, which you specifically reference in your rationale on p. 19. Why is this aspect omitted?

Question for developer: The proposed measure does not include a question to assess whether the patient was asked about his/her goals or preferences for treatment? The current questions talk about reasons to have or not have the intervention, but this is more related to pro/con and alternatives discussions than patient goals for treatment of their condition. Does this process measure need to include such assessment to represent SDM fully?

**In general, there is evidence of a performance gap (room for practices to grow), but this may vary by condition (or medical sub-specialty) as indicated by the higher SDM scores for patients considering surgery for lower back pain. Accordingly, the proposed measure may be of greater value for some specific conditions.

1c. PRO-PM
Comments:
**Developer references patient surveys about the value patients place on being informed and included in decision making about treatment options. Example: Wexler et al 2015 noted broad support for importance of decision aids to patient decision making. Survey data from Dartmouth-Hitchcock Med Center and Public Opinion Strategies also conclude that patients believe they have an equal if not primary role in decision making and that decision aids and discussion with the clinician are important to

make appropriate treatment choices.

**The developers describe studies that show that patients value SDM as a global concept. More evidence is needed to support the content validity (meaningfulness and importance) of the 4 items included in the measure.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. **Data source(s):** administrative claims, questionnaire (4 questions)/patient reported data/survey **Specifications:** • The numerator is the sum of numerical assigned values to question responses; the questions and instructions on how to assign the numerical score are clearly defined. • The denominator is calculated via patient identification via administrative claims. • ICD-10 and CPT codes are provided for the 7 conditions and for exclusions, where required. • This measure is not risk-adjusted. ***Questions for the Committee :*** ○ *Do you have any questions on the specifications, codes, definitions, etc.* ○ *Are all the data elements clearly defined? Are all appropriate codes included?* ○ *Is the logic or calculation algorithm clear?* ○ *Is it likely this measure can be consistently implemented?*
2a2. Reliability Testing Testing attachment
2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

The following data sources were used for testing this measure:

1) TRENDS, a national survey of adults over 40 in Knowledge Networks panels who had made decisions were used to estimate "usual care" experience for back, knee and hip decision experience (2012)
2) Surveys of Medicare patients who had mastectomy, prostate cancer surgery or PCI were used to estimate usual care experiences for those procedures (2008)
3) Demonstration site data. Nearly 3000 patients were surveyed in 6 different clinical sites around the US that were implementing the use of decision aids and encouraging shared decision making for 14 different decisions about testing and surgery.
4) These data were collected from 2009 through 2013.

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure

☒ **Yes** ☐ **No**

Method(s) of reliability testing

The developer indicated the following types of testing:

1. At the item level, we measured test-retest reliability from same individuals 4 weeks apart
2. At the item and score levels for an encounter, we compared patient reports with coding of tape recordings of encounters
3. At the practice level, we randomly split patients making the same decision at the same clinical site into groups of 25 or larger and correlated the scores; i.e. how well score from one sample's reports correlated with another sample's reports for same decision for same provider group.

Results of reliability testing The developer provide the following summary of reliability testing results:

- The developer reported short term (~4 weeks) test-retest data on some variations of this measure and obtained ICC values in the .7 to .8 range.
- The developer also provided two tests of whether or not patient reports of their interactions align with coding of tape recordings of the same interactions. In one study, objective observers and patients both rated various aspects of the interactions between doctors and patients making breast cancer decisions. The results showed a high level of agreement, although patients' ratings tended to a bit higher, on average, than observers'(Pass et al, 2012) In a different test, women's interactions with physicians about primary treatment for breast cancer were tape recorded (n = 96). Coding of the interactions were related to patient reports using the questions in the Process Score. In this case, because the clinically reasonable options were known, questions were asked separately for discussion of the pros and cons of both reasonable options. Kappas were computed for the dichotomous variables and product moment correlations for the multi-category items between the coded results and what respondents said. For the overall scores, the correlations were .50 ($p<.001$) and .38 ($p=.004$) for adjuvant therapy and surgery decisions respectively. With respect to individual items, the values were higher for whether the options were presented (.64 to .71) and how much the reasons for each option were discussed (.64 to .75) and lower for how much the cons were discussed (.16 to .46) and whether the patient's input was sought (.14 to .32).
- Finally, for reliability at the level of clinical practice, the developer divided patients from the same site making the same decision into random groups and correlated their Process Scores.

With minimum sample sizes of 25, there was an average reliability of .61. The developer noted an expectation that the numbers would be higher with larger samples.

Guidance from the Reliability Algorithm
Submitted Specifications precise, unambiguous and complete (Box 1)Yes →Empirical reliability testing conducted on measure as specified (Box 2) Yes → Reliability testing with computed performance measure score (Box 4) Yes →Method was appropriate (Box 5) Yes → Based on results, what level of certainty (high, moderate, low) or confidence that performance measure scores are reliable (Box 6) → Moderate.

- Staff rating of moderate due to the <.7 average from the clinical practice scores.

Questions for the Committee:

- *Are the tests sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ Yes ☐ Somewhat ☐ No
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ Measure score ☐ Data element testing against a gold standard ☐ Both

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
The developer indicates the following: *The evidence for the value of clinical practices devoted to shared decision making and that the SDP score is a valid measure of clinical performance comes from a number of studies of decision making in clinical practices, some of which were trying to implement*

shared decision making on a routine basis and using decision aids for many decisions.

- The developers compared the aggregate Process Scores from patients treated at clinical sites that have committed to shared decision making, usually by including the routine use of decision aids, with reports of national cross-sections of patients from the TRENDS survey who made the same decisions.
- The developer indicated that a better test may come from studies of breast cancer decision making in four clinical sites. One of these four sites routinely used decision aids and had support for patients when they met with their surgeons to facilitate getting patients' questions asked and answered. The other three sites practiced usual care, with no special intervention to encourage shared decision making. Similar data is available for decision making around hip and knee replacement.
- Finally, a small study at a clinical site in Stillwater, Minnesota collected data using the Decision Process Score questions from patients who discussed treatment for benign prostatic hyperplasia (BPH) with their urologists. They started collecting these data before introducing decision aids and continued to collect them after the use of decision aids that encouraged shared decision making became routine in the practice.

Validity testing results:

- Refer to the tables in the testing attachments
- The developers summarized: W*e have data that show clearly that decision making on average in the US as measured by this score is not very good and that clinical sites that commit to improved decision making attain average scores from their patients that are much higher than average. We think this is one of relatively few instances in which outcome measures based on patient reports are clearly linked to the way that clinical practices are trying to interact with patients.*

Questions for the Committee:

- *Are the test samples adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The exclusions for these measures only apply to two of the decisions.
  - PCI patients who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).
  - For patients who have mastectomy, patients who had had a prior lumpectomy for breast cancer in the same breast and patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies) are excluded.
- The developer indicates: the exclusions are specifically targeted to focus on those patients for whom shared decision making is clearly appropriate.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? Note: Since this information is not provided – do you think it is necessary to inform your rating of the validity?*

2b4. Risk adjustment: Risk-adjustment method ☒ None ☐ Statistical model ☐ Stratification

Risk adjustment summary The developer provides the following rationale for lack of risk adjustment:

- There are two reasons we are not recommending any kind of risk adjustment. First, and perhaps most important, there is no ethical basis for saying the standards for engaging in shared decision making for these preference-sensitive surgical decisions should vary by patient characteristics. Second, as the data on disparities above shows, we have not found any systematic differences in average decision making scores based on age, gender, education or ethnicity. So, in our experience, adjustments would not have any meaningful effect on results.

Questions for the Committee:

- *If a justification for no risk adjustment is provided, is there any evidence that contradicts the developer's rationale and analysis?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developers used t-tests to assess differences between mean Shared Decision Process scores from patient who made decisions in practices that had implemented procedures to promote shared decision making and patients who made decisions in usual care, either based on data from practices or from our national surveys.
- The practices using decision aids and promoting shared decision making consistently had significantly better scores on this measure. The exception is decisions about surgery for lower back pain, which consistently get very high scores in "usual care". We think that is not a reflection of a problem with the measure but a reflection of the way back surgery decisions are made.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

Developer information: it is relatively rare for respondents to the follow-surveys not to answer all four questions.

- For example, from our test sites from patients making decisions about knee replacement and lower back surgery for herniated disc and spinal stenosis, the percentages of respondents not answering all four questions were 3%, 5% and 0% respectively; the percentages having more than one missing response were <1%, 2% and 0% respectively, from a total of 411 respondents.
- We did not experiment with alternative ways of handling missing data, because it really could not affect the results. We think leaving out anyone not answering all the questions or imputing a .5 score for one missing response (and eliminating anyone with more than one missing answer) would both be reasonable approaches to dealing with missing data.

Guidance from the Validity Algorithm

Measure specifications consistent with evidence (Box 1); Yes →Potential threats to validity assessed (Box 2)→provided rationale where appropriate (missing data, exclusions, lack of risk adjustment) → Empirical testing conducted using measure as specified (Box 3); Yes →Validity testing conducted with computed performance scores for measured entity (Box 6); Yes → Method Described Appropriate (Box 7); Yes →Based on results, scope and analysis of testing, level (high, moderate) of certainty or

confidence that the measure is a valid indicator of quality (Box 8); Moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**Specifications for the measure questionnaire and calculation of the aggregate SDM process score are clearly defined. The developer notes no need to risk adjust the measure. Survey and sampling instructions clearly delineate that the measure would only calculate scores for patients electing one of 7 surgical procedures.

Questions for Developer: Will time lag between patient election of a surgical procedure and actual post-discharge administration of the survey affect reliability? What is your expectation regarding patient response rates and will that affect the value of a score (e.g., does it matter whether the score for a given practice reflects one patient vs. 25 patients)?

2b.1 - The specifications are consistent with what the target population values in terms of sharing of information about options for treatment and inclusion in decision making about the procedure.
Questions for developer: The proposed process measure does not include a question to assess presence or use of decision aids, which you specifically reference in your rationale on p. 19. Why is this aspect omitted?

Question for developer: The proposed measure does not include a question to assess whether the patient was asked about his/her goals or preferences for treatment? The current questions talk about reasons to have or not have the intervention, but this is more related to pro/con and alternatives discussions than patient goals for treatment of their condition. Does this process measure need to include such assessment to represent SDM fully?

**There is need for more precise specifications:

Developers note that meaningful comparisons of PM scores across sites should only for a particular decision. How are these decisions defined? Should patients that are making multiple decisions be excluded?

Developers suggest 2 alternatives for dealing with missing data. A single recommendation is needed to avoid biasing organization's PM scores.

2a2. Reliability Testing

Comments:

**How is the measure being publicly reported?

The developer does not indicated the intent to publicly report this measure – rather, it is intended as an internal measure associated with certification (e.g., Quality Path Program sponsored by Alliance) or evaluation against stated strategic goals for accountability and patient engagement.

For maintenance measures – which accountability applications is the measure being used for? N/A

How can the performance results be used to further the goal of high-quality, efficient healthcare?
Effective use of a PRO measure on SDM process can give further evidence to the value and impact of SDM, to identifying aspects of the SDM process that have impact from a patient perspective and to offer feedback to a clinical practice about its SDM practices and potential areas for improvement.

Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them?
The cost and administrative burden of this measure is not known or quantified by the developer and information should be supplied to NQF.

The developer elects to focus on surgical procedure rather than assessment of shared decision making without reference to elected procedure. This assumes that for the conditions targeted, only surgical interventions are or should be part of SDM. One unintended consequence is that practices focus on SDM only for surgical candidate patients, missing an opportunity to engage in SDM with a patient group with a specific diagnosis to assess a full range of treatment options. While one can see a benefit in creating a method for quantifying value of SDM as a measure of quality, it's not clear that this proposed measure will do so in a way this a fully meaningful for patients facing decisions within these disease states.

**Clearer rationale for exclusion of structural analyses (e.g., factor) and estimation of internal consistency reliability are needed.

Test-retest reliability is moderately strong.

Reliability at the practice level is barely adequate (ICC = .61), but the authors note that they may have insufficient sample sizes. Additional evidence of practice-level reliability should be provided as it becomes available.

2b.2 Validity Testing
Comments:
**What level was tested : Performance Measure Score using empirical validity testing
Was reliability tested with an adequate scope (number of entities and patients) to generalize for widespread implementation; AND with an appropriate method?

The developer presents validity testing data comparing demo sites with scores from national TRENDS survey of patients making the same decision, comparing demonstration sites with “usual care” sites where SDM not in practice, and within a practice before and after the use of decision aids occurred. Results consistently showed higher composite scores for demo site patients and after introduction of decision aid use.

Describe how the results either do or do not demonstrate sufficient validity so that conclusions about quality can be made?
Results appear to demonstrate validity of the SDM process measure as a marker of quality, as expressed by higher score by patients for the clinical action of discussion and inclusion in determining best treatment course. The results support conclusions that clinical practitioners that engage in behaviors to stimulate discussion and shared decision making will garner more positive patient reports for inclusivity and more comprehensive information shared with patients.

**PM score differences between clinics that apply "usual care" vs. decision aids provides compelling evidence of the measure's validity at the practice level.

At the individual respondent level, there is need for greater evidence of the measure's content validity. The developers note that the item development was informed by "extensive cognitive testing," but these procedures were not described.

2b3.-2b7. Test Related to Potential Threats

Comments:

**Are exclusions supported by the evidence and/or analyses that indicate sufficient frequency?
Yes

2b4.If outcome, PRO, or resource use performance measure: Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results?

This measure is not risk adjusted. The developers addressed this through exclusions, to focus the measure on instances when alternatives for treatment intervention exist. Exclusion of elective mastectomy and prior recent cardiac arrest or previous coronary interventions were deemed appropriate to focus on true instances of shared decision making opportunity.

2b5. How do analyses indicate this measure identifies meaningful differences about quality?

The developer notes that practices using decision aids and promoting SDM generally present higher scores on similar PRO measure in limited use. The use of this SDM process measure is assumed to be a proxy for quality in that SDM is deemed a factor in the delivery of quality clinical care. The measure will ostensibly identify practices and procedures for which SDM is occurring with higher frequency, higher overall satisfaction based on the composite score on the survey.

2b6. If multiple sets of specifications: Do analyses indicate they produce comparable results? N/A

2b7. If eMeasure, composite, PRO-PM: Do analyses indicate missing data does not bias results?
The developers do not address missing data impact on results but concede that the easiest way to handle this possibility is omission of the patient response when at least one question is unanswered.

**Sampling bias is the greatest threat to validity of the proposed measure. What percentage of potential respondents returned a completed survey? This is a substantial concern since the percentage of respondents may differ across sites and influence PM scores.

It is notable that the psychometric analyses were conducted with sample that is 96% white and 63% male.

Older patients are also disproportionately represented, but this may reflect the conditions under evaluation.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Data are collected via patient survey.
- The method used most often is to send a questionnaire by mail to eligible respondents with a postage paid return envelope. A second mailing and a telephone reminder have been used to increase the rate of response. There may be settings in which responses can be collected via the Internet.
- The developer notes that as long as responses are self-administered rather than interviewer administered, we think that is fine. However, data collection protocols must make it clear that individual answers will not be viewed by the physician and/or his/her staff. Results from similar surveys have made it clear that survey responses are skewed if respondents think they can be reviewed by their providers or their support staffs. Therefore, we recommend that data not be collected from respondents who are in a clinic or hospital setting.
- The questions are publicly available with no fees for the survey.
- Approximate fees/expenses incurred for administration of the survey by a clinical practice were not provided.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*
- *Do you think a clinical practice would find the implementation of this measure feasible, without excessive burden?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3 Feasibility

Comments:

**In most practices, data elements required for this process PRO measure are not generated during care delivery. Patients are not routinely asked whether decision aids were shared, whether pros/cons were discussed or whether alternative interventions were presented for their consideration. This proposed measure introduces a methodology that takes place outside of and after care delivery.

Which of the required data elements are not available in electronic form, e.g., EHR or other electronic sources?

None of the information solicited by this measure are EHR/electronic source. Developer indicates survey is paper based and mailed to patient, identified by their election of one of the 7 surgeries, post discharge from having the procedure.

What are your concerns about how the data collection strategy can be put into operational use?

I have concerns about the impact of timing and method (mailed questionnaire with/without phone followup) on patient response rate. The developer notes that conducting such a survey at discharge or during inpatient stay may skew results because patients may fear retaliation or lack of anonymity to clinicians and care takers. The developer even notes that the questionnaire should explicitly note that

the clinician will not see their individual answers. They offer no evidence of such bias in similar survey strategies. The proposed approach presents both an operational hassle (expensive followup is assumed but no detail is provided and reliability of data based on type and intensity of followup) and potential for gaps in recall (decisions may be made weeks before a procedure) or influence of results based on surgical outcome (e.g., will complications from surgery delay response or skew response if patient perceives negative experience?)

Question for developer: What are the cost and administrative burden estimates associated with this measure?

**Developers note that response rates improved with telephone follow-up, which may not be feasible for some practices.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
The developer states "The Foundation is not in position to sponsor or implement quality accountability measurement. However, we think the efforts described in 4.1 are examples of the kinds of programs that will want to start using this measure when it is NQF approved."

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**

OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

Improvement results New Measure

Potential harms None identified

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☐**X Moderate** ☐ **Low** ☐**Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4 Usability and Use

Comments:
**How is the measure being publicly reported?

The developer does not indicated the intent to publicly report this measure – rather, it is intended as an internal measure associated with certification (e.g., Quality Path Program sponsored by Alliance) or evaluation against stated strategic goals for accountability and patient engagement.

For maintenance measures – which accountability applications is the measure being used for? N/A

How can the performance results be used to further the goal of high-quality, efficient healthcare?
Effective use of a PRO measure on SDM process can give further evidence to the value and impact of SDM, to identifying aspects of the SDM process that have impact from a patient perspective and to offer feedback to a clinical practice about its SDM practices and potential areas for improvement.

Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them?
The cost and administrative burden of this measure is not known or quantified by the developer and information should be supplied to NQF.

The developer elects to focus on surgical procedure rather than assessment of shared decision making without reference to elected procedure. This assumes that for the conditions targeted, only surgical interventions are or should be part of SDM. One unintended consequence is that practices focus on SDM only for surgical candidate patients, missing an opportunity to engage in SDM with a patient group with a specific diagnosis to assess a full range of treatment options. While one can see a benefit in creating a method for quantifying value of SDM as a measure of quality, it's not clear that this proposed measure will do so in a way this a fully meaningful for patients facing decisions within these disease states.

Criterion 5: Related and Competing Measures

Related or competing measures

1741 : Patient Experience with Surgical Care Based on the Consumer Assessment of Healthcare Providers and Systems (CAHPS)® Surgical Care Survey

Harmonization

The approved PCMH and ACO CAHPS measures of shared decision making were adaptations of the measures we developed and are proposing. Those measures were used for respondents who reported they had discussed starting or stopping a prescription medication (for PCMH) and for patients who reported discussion a prescription medication or a procedure with a provider (ACO). The problem with integrating this measure into the CAHPS protocols includes both sample sizes and sample designs. This measure works best when applied to a specific kind of decision (eg. Decision to take medication for high blood pressure or decision to have surgery for herniated disc.) CAHPS samples relatively small numbers of ambulatory patients from a clinician's practice or a clinical site. Those samples do not include enough encounters at which decisions are made about specific medications or specific tests or surgical procedures to provide reliable data. Hence, they had to ask about any decisions about starting or stopping medications or surgical procedures and combine the answers for each type of decision. The numbers of such decisions tend to be very small, even when all medications or procedures are combined. Moreover, we have abundant data showing that the Shared Decision Making

Process Score varies widely from medication to medication and procedure to procedure. (Zikmund=Fisher et al, 2010; Fowler et al, 2012; Fowler et al, 2014). The approach we are proposing, sampling patients who have undergone a procedure, provides the ability to control the sample sizes of respondents and provides for collecting data about the same decision when using the data to compare clinical sites—which is essential in order to meaningfully interpret the results as measures of quality of care.

Pre-meeting public and member comments

Name: Ms. Suzanne Pope

Organization: American Urological Association

Comment: For consideration: should this measure also include patients who have radiation therapy for prostate cancer (i.e., why is SDM critical only for radical prostatectomy among the treatment options? What about active surveillance? It would seem that a more inclusive measure would be to measure SDM agnostic to what option was chosen.)

Name: Megan Burke, MSW

Organization: The SCAN Foundation

Comment: The SCAN Foundation acknowledges the importance of shared decision-making as part of person and family-centered care (PFCC). The proposed measures capture the time a doctor spent discussing pros and cons of a procedure, and the individual's choices. However, PFCC quality measures should also assess whether the provider elicited information from the individual about his/her goals, and discussed how treatments do or do not align with the stated goals.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Shared Decision Making Process
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: Click here to enter a date

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence*

form to the individual measure submission.

- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☒Patient-reported outcome (PRO): Appropriate shared decision making process when decisions are made about any one of 7 surgical procedures for which there are reasonable alternatives.
PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

When faced with a medical problem for which there is more than one reasonable approach to treatment or management, shared decision making means providers should outline for patients that there is a choice to be made, discuss the pros and cons of the options and make sure that patients have input into the final decision. The result will be decisions that align better with patient goals, concerns and preferences. This measure asks patients who had any of 7 preference sensitive surgical interventions to report on the interactions they had with their providers when the decision was made to have the surgery.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

When physicians provide balanced information to patients (often in the form of decision aids) and have a discussion about the options and about what patients want, patients answers these questions in a way that reflects a shared decision making process.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2962
De.2. **Measure Title:** Shared Decision Making Process
Co.1.1. **Measure Steward:** Informed Medical Decisions Foundation, a division of Healthwise
De.3. **Brief Description of Measure:** This measure assesses the extent to which health care providers actually involve patients in a decision-making process when there is more than one reasonable option. This proposal is to focus on patients who have undergone any one of 7 common, important surgical procedures: total replacement of the knee or hip, lower back surgery for spinal stenosis of herniated disc, radical prostatectomy for prostate cancer, mastectomy for early stage breast cancer or percutaneous coronary intervention (PCI) for stable angina. Patients answer four questions (scored 0 to 4) about their interactions with providers about the decision to have the procedure, and the measure of the extent to which a provider or provider group is practicing shared decision making for a particular procedure is the average score from their responding patients who had the procedure.
1b.1. **Developer Rationale:** We have collected a great deal of data from cross-section surveys of patients who have made decisions and from patients drawn from clinical sites documenting that for many decisions, patients routinely do not perceive that they discuss the cons of proposed interventions, are not told about alternatives and are not asked for their input into the decisions. Consistently, their levels of knowledge of information relevant to the decisions they are making are low. We then have evidence that when clinicians commit to shared decision making, by routinely providing decision aids for example, the scores of patients with respect to knowledge and the decision making process are higher. We believe the use of the Shared Decision Making Process Score and appropriate measures of patient knowledge can be catalysts to routinely informing and involving patients in important medical decisions, which in turn will increase the likelihood that patients will get the care they want and that is consistent with their goals and concerns (Stacey et al 2014).

S.4. **Numerator Statement:** Patient answers to four questions about whether not 4 essential elements of shared decision making (laying out options, discussing the reasons to have the intervention and not to have the intervention, and asking for patient input) were part of the interactions with providers when the decision was made to have the procedure.
S.7. **Denominator Statement:** All responding patients who have undergone one of the following 7 surgical procedures: back surgery for a herniated disc; back surgery for spinal stenosis; knee replacement for osteoarthritis of the knee; hip replacement for osteoarthritis of the hip; radical prostatectomy for prostate cancer; percutaneous coronary intervention (PCI) for stable angina, and mastectomy for early stage breast cancer.
S.10. **Denominator Exclusions:** For back, hip, knee, and prostate surgery patients, there are no exclusions, so long as the surgery is for the designated condition.

PCI patients who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).

For patients who have mastectomy, patients who had had a prior lumpectomy for breast cancer in the same breast and patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies) are excluded.

De.1. **Measure Type:** PRO
S.23. **Data Source:** Patient Reported Data/Survey
S.26. **Level of Analysis:** Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_evidence-635947722045099247.docx,NQF_table_attachments.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
We have collected a great deal of data from cross-section surveys of patients who have made decisions and from patients drawn from clinical sites documenting that for many decisions, patients routinely do not perceive that they discuss the cons of proposed interventions, are not told about alternatives and are not asked for their input into the decisions. Consistently, their levels of knowledge of information relevant to the decisions they are making are low. We then have evidence that when clinicians commit to shared decision making, by routinely providing decision aids for example, the scores of patients with respect to knowledge and the decision making process are higher. We believe the use of the Shared Decision Making Process Score and appropriate measures of patient knowledge can be catalysts to routinely informing and involving patients in important medical decisions, which in turn will increase the likelihood that patients will get the care they want and that is consistent with their goals and concerns (Stacey et al 2014).

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See answer to 1b.3 below

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
We have data comparing clinical sites that have made a commitment to do shared decision making with the shared decision making process scores in usual care, derived both from cross-section surveys of patients who have made the decisions or clinical sites that were making no special effort to implement shared decision making. The data in the five tables in the attachment consistently show that clinical sites that made a special effort to implement shared decision making have "significantly" higher shared decision making process scores from their patients than patients in "usual care".
These data are presented in detail in NQF table attachment

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

We have data comparing reported shared decision making process scores by patient age, gender, education and race. Although there are some examples of significant relationships in the data, they do not go in consistent directions. The takeaway from the data is that we do not have evidence that the processes of decision making with providers are consistently related to any of those patient demographic characteristics.
These results are presented in detail in NQF table attachment

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
See 1b.4

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
These 7 procedures are very prevalent, as shown by the number of procedures covered by Medicare. In addition, there is wide geographic variation in the rates at which these procedures are performed, which is widely interpreted as evidence that these are provider-driven decisions that would benefit from more patient knowledge and involvement.
These data are presented in NQF table attachment

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
http://www.dartmouthatlas.org/tools/downloads.aspx?tab=41

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Every time we have surveyed patients about their support for the use of decision aids and how they want to be involved in medical decision making, the results have been overwhelming support for both. For example, data from 2800 patients who were given decision aids when faced with medical decisions found that 83% said that it was “extremely” or “very” important that patients receive decision aids when making decisions like the ones they made. (Wexler,et al., 2015).

Data from patients at Dartmouth-Hitchcock Medical Center elaborate on these research findings. This is one of the sites where before making important medical decisions, patients routinely receive decision aids and fill out post-viewing questionnaires. Analysis of questionnaire responses relating to fourteen different decisions regarding topics ranging from prostate cancer screening with the PSA test to back surgery, substantiated patients’ interest in shared decision making. When asked who should make the decision (“mainly me,” “mainly the doctor,” or “both equally”), a majority said “mainly me” for all but two of the decisions, and more than 90 percent said either “mainly me” or “both equally” for every one of the 14 decisions.

A recent cross-sectional survey of adults, which was conducted by Public Opinion Strategies, provides additional evidence that patients want to be involved in decision making. Respondents were asked to read a statement about informed decision making (shown below) and rate their favorability toward the concept on a scale from 0 to 100. “Informed medical decision making is an idea in health care that patients should receive information about all of the treatment choices and options available to them for a specific disease, illness, or procedure before they decide, in conjunction with their doctor, on the appropriate treatment choices.” With 100 being the most favorable response, the mean rating was 82. Almost 70 percent of respondents rated the statement with a score greater than 80. These data show that when given clearly worded questions about medical decision making, the majority of people want to be involved in an active decision making process.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer : Breast, Cancer : Prostate, Cardiovascular : Percutaneous Coronary Intervention (PCI), Musculoskeletal : Joint Surgery, Musculoskeletal : Low Back Pain, Surgery : Cardiac Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.massgeneral.org/decisionsciences/research/DQ_Instrument_List.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ICD_Codes.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patient answers to four questions about whether not 4 essential elements of shared decision making (laying out options, discussing the reasons to have the intervention and not to have the intervention, and asking for patient input) were part of the interactions with providers when the decision was made to have the procedure.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The data should be collected from a sample of patients of a target provider or provider group as soon as possible after the procedure is performed, not more than 6 months afteer the procedure was performed with no fewer than 50 respondents. This submission takes no position on how often a provider or provider group should be evaluated.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome

should be described in the calculation algorithm.
All responding patients will answer four questions about their pre-surgical interactions with their providers:

1. How much did a doctor (or health care provider) talk with you about the reasons you might want to (HAVE INTERVENTION)—a lot, some, a little, or not at all?
2. How much did a doctor (or other health care provider) talk with you about reasons you might not want to (HAVE INTERVENTION)—a lot, some, a little or not at all?
3. Did any of your doctors ask you if you wanted to (HAVE INTERVENTION)? (YES/NO)
4. Did any of your doctors (or health care providers) explain that you could choose whether or not to (HAVE INTERVENTION)? (YES/NO)

OR: "Did any of your doctors (or health care providers) explain that there were choices in what you could do to treat your [condition]? (YES/NO)

SCORING: 1 POINT EACH FOR ANSWERING "A LOT" OR "SOME" TO QUESTIONS 1 AND 2; 1 POINT EACH FOR ANSWERING "YES" TO QUESTIONS 3 AND 4. TOTAL SCORE = 0 TO 4.

Score for a provider or provider group is simply the average score for their responding patients. This will be a continuous number from 0 to 4.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All responding patients who have undergone one of the following 7 surgical procedures: back surgery for a herniated disc; back surgery for spinal stenosis; knee replacement for osteoarthritis of the knee; hip replacement for osteoarthritis of the hip; radical prostatectomy for prostate cancer; percutaneous coronary intervention (PCI) for stable angina, and mastectomy for early stage breast cancer.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See S2. There is an attached sheet with ICD 10 and CPT codes needed to identify eligible patients.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
For back, hip, knee, and prostate surgery patients, there are no exclusions, so long as the surgery is for the designated condition.

PCI patients who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).

For patients who have mastectomy, patients who had had a prior lumpectomy for breast cancer in the same breast and patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies) are excluded.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Included in attached file

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
none

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Continuous variable, e.g. average
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
All responding patients will answer four questions about their pre-surgical interactions with their providers:

1. How much did a doctor (or health care provider) talk with you about the reasons you might want to (HAVE INTERVENTION)—a lot, some, a little, or not at all?
2. How much did a doctor (or other health care provider) talk with you about reasons you might not want to (HAVE INTERVENTION)—a lot, some, a little or not at all?
3. Did any of your doctors ask you if you wanted to (HAVE INTERVENTION)? (YES/NO)

Did any of your doctors (or health care providers) explain that you could choose whether or not to (HAVE INTERVENTION)? (YES/NO)
OR: "Did any of your doctors (or health care providers) explain that there were choices in what you could do to treat your [condition]? (YES/NO)

SCORING: 1 POINT EACH FOR ANSWERING "A LOT" OR "SOME" TO QUESTIONS 1 AND 2; 1 POINT EACH FOR ANSWERING "YES" TO QUESTIONS 3 AND 4. TOTAL SCORE = 0 TO 4.

Score for a provider or provider group is simply the average score for their responding patients. This will be a continuous number from 0 to 4.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Patients of a particular surgeon or at a particular clinical site (which could be a group of providers or a hospital or other surgical site) who had one of the 7 target procedures are identified from medical records, claims or in some other way. Patients can be sampled sequentially, or a pool of such patients who had the procedure in a particular time period can be created and sampled at a rate that produces the desired number of potential respondents.

These same questions can be used in a population-based sample, such as a sample of a population in a geographic area. Eligible

respondents could be identified from claims (such as Medicare claims files) or based on patient self- reports of having had the procedures within some time frame. The measures have been used in surveys using both of those models. However, the basic proposal here is to use the measure to evaluate clinical care provided by particular clinical sites, provider groups, or providers.

With respect to sample sizes, the standard deviations vary some by procedure. For most procedures, comparing samples of size of 50 or larger will detect differences of .5 in Decision Process Scores ($p < .05$), which is an order of magnitude we have often observed between sites that do and do not make an effort to do shared decision making. Samples of 100 reduce that number to around .3. We think samples in the range of 50 to 100 offer sufficient power to detect clinically meaningful differences in clinical practice.

IF a PRO-PM, identify whether (and how) proxy responses are allowed.

Proxy respondents are not permitted. Virtually all of the patients who receive these procedures should be able to answer survey questions. We think it is important to get the perceptions of the patients themselves about the process.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
The method we have used most often is to send a questionnaire by mail to eligible respondents with a postage paid return envelope. A second mailing and a telephone reminder have been used to increase the rate of response. There may be settings in which responses can be collected via the Internet. So long as responses are self-administered rather than interviewer administered, we think that is fine. However, data collection protocols must make it clear that individual answers will not be viewed by the physician and/or his/her staff. Results from similar surveys have made it clear that survey responses are skewed if respondents think they can be reviewed by their providers or their support staffs. Therefore, we recommend that data not be collected from respondents who are in a clinic or hospital setting.

IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

We recommend that data not be accepted if response rates are lower than 50%. Calculate response rate as all those responding divided by all those invited to answer the survey questions (AAPOR response rate 4).

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
If one answer is missing, assign value of .5. If more than one answer is missing, delete the case.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Patient Reported Data/Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
We have used these questions in mail surveys most often, but we have also use them on the Internet and in a national telephone survey using telephone interviewers. We have used these questions in English and Spanish.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_attachment_fowler_final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Shared Decision Making Process
Date of Submission: Click here to enter a date
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance**

measures, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of

exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Patient reports Click here to describe	☒ other:Patient reports Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
1) TRENDS, a national survey of adults over 40 in Knowledge Networks panel who had made decisions were used to estimate "usual care" experience for back, knee and hip decision experience (2012)
2) Surveys of Medicare patients who had mastectomy, prostate cancer surgery or PCI were used to estimate usual care experiences for those procedures (2008)
3) Demonstration site data. Nearly 3000 patients were surveyed in 6 different clinical sites around the US that were implementing the use of decision aids and encouraging shared decision making for 14 different decisions about testing and surgery. These data were collected from 2009 through 2013. The numbers varied by procedure and are presented in the appropriate tables. These data were used to assess the decision making process scores for patients in setting in which clinical sites were making an effort to implement shared decision making.

They were also used to estimate the reliability of average Shared Decision Making Process scores for clinical sites. Most of the usable data for that analysis came from Dartmouth medical center, because they had the most

responses, and we wanted 20 or more responses in each random half estimate of the rating at a practice for a particular decision. The analysis was based on responses from 663 patients over 5 different decisions.

In addition, we had data from 4 clinical sites from 266 patients who made decision about breast cancer treatment, 1 site emphasizing use of DAs and shared decision making and the other three in "usual care" mode which we used to add to our validity data.

1.3. What are the dates of the data used in testing? 2008 to 2014 Click here to enter date range

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Although we used our cross-sectional data from surveys for estimates of usual care values, including means and SDs, all of the evidence for the values related to the validity and reliability of the measure to reflect clinical practice represents average patient reported scores, either for individual practices or a combination of practices that either were or were not making a special effort to promote shared decision making

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The cooperating practices all varied in the decisions for which they used decision aids and how they were distributed. The data on the Shared Decision Making Process in demonstration sites were mainly collected by sending out a mail questionnaire after patients had met with providers about the decisions. Response rates varied by site and decision. We usually included all patients who completed a questionnaire.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when

SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Profile of all patients respondent in the demonstrations sites:

Respondent characteristics

Characteristic[a]	Respondents n (%)
Age group (n = 2,928)	
<50	438 (15)
50 - 64	1,533 (52)
≥ 65	957 (33)
Gender (n = 2,961)	
Male	1,881 (63)
Female	1,080 (37)
Education (n = 2,914)	
High school or less	926 (32)
Some college or 2-y college	819 (28)
4-year college or more	1,169 (40)
Race (n = 2,832)	
White	2,721 (96)
Black	68 (2)
Ethnicity: (n = 2,893)	Hispanic 62 (2)

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

1. At the item level, we measured test-retest reliability from same individuals 4 weeks apart
2. At the item and score levels for an encounter, we compared patient reports with coding of tape recordings of encounters
3. At the practice level, we randomly split patients making the same decision at the same clinical site into groups of 25 or larger and correlated the scores; i.e. how well score from one sample's reports correlated with another sample's reports for same decision for same provider group.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The Decision Process Score is technically a composite, with conceptual roots in what a good decision process should look like, so a calculation of Cronbach's alpha may not be an appropriate measure of reliability (see Bollen and Lennox, 1991)[1], but we have calculated them for some decisions, and they are reasonably high (often in the .5 to .7 range)

We have short term (~4 weeks) test-retest data on some variations of this measure and obtained ICC values in the .7 to .8 range.

We also have two tests of whether or not patient reports of their interactions align with coding of tape recordings of the same interactions. In one study, objective observers and patients both rated various aspects of the interactions between doctors and patients making breast cancer decisions. The results showed a high level of agreement, although patients' ratings tended to a bit higher, on average, than observers'(Pass et al, 2012)[2].

In a different test, women's interactions with physicians about primary treatment for breast cancer were tape recorded (n = 96). Coding of the interactions were related to patient reports using the questions in the Process Score. In this case, because the clinically reasonable options were known, questions were asked separately for discussion of the pros and cons of both reasonable options. Kappas were computed for the dichotomous variables and product moment correlations for the multi-category items between the coded results and what respondents said. For the overall scores, the correlations were .50 (p<.001) and .38 (p=.004) for adjuvant therapy and surgery decisions respectively. With respect to individual items, the values were higher for whether the options were presented (.64 to .71) and how much the reasons for each option were discussed (.64 to .75) and lower for how much the cons were discussed (.16 to .46) and whether the patient's input was sought (.14 to .32).

Finally, for reliability at the level of clinical practice, we have divided patients from the same site making the same decision into random groups and correlated their Process Scores. With minimum sample sizes of 25, we get an average reliability of .61. The numbers would be higher with larger samples, which we hope to have soon.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., *what do the results mean and what are the norms for the test conducted?*)

__We think the reliability of the overall process is satisfactory at both the individual encounter level and at the clinical practice level. In particular, at the practice level, which is the level that is more relevant for the way we propose to use this measure, we think the reliability will only get higher with bigger samples.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
We have a number of ways we have looked at validity. The approach is described below with each individual approach to testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The evidence for the value of clinical practices devoted to shared decision making and that the SDP score is a valid measure of clinical performance comes from a number of studies of decision making in clinical practices, some of which were trying to implement shared decision making on a routine basis and using decision aids for many decisions. The following summarizes those results.

We have compared the aggregate Process Scores from patients treated a clinical sites that have committed to shared decision making, usually by including the routine use of decision aids, with reports of national cross-sections of patients from the TRENDS survey who made the same decisions.

Table 2. Mean Decision Process Scores at SDP Demonstration sites and from a national sample of patients for three orthopedic procedures.

Data Source	Decision Topic	N	Mean Process Score	Std. Deviation
TRENDS	Surgery: Knee Pain	163	2.81	1.139
Demo Sites	Knee Osteoarthritis	239	3.24**	.840
TRENDS	Surgery: Hip Pain	57	2.45	1.236
Demo Sites	Hip Osteoarthritis	129	3.31***	.864
TRENDS	Surgery: Low Back Pain	152	3.23	1.016
Demo Sites	Herniated Disc + Spinal Stenosis	55	3.38	.828

**p < .01

*** P< .001

For osteoarthritis of the knee and hip, it can be seen that the patients in practices where decision aids are used reported significantly better decision processes than a cross-section sample of adults who faced the same decisions. The responses did not differ for conversations about lower back pain, but the decisions about back pain were by far the best decision processes based on respondent reports in the national survey.

Because the data in the above table were collected with quite different time periods between the decision and the measurement, a better test may come from studies of breast cancer decision making in four clinical sites. One of these four sites routinely used decision aids and had support for patients when they met with their surgeons to facilitate getting patients' questions asked and answered. The other three sites practiced usual care, with no special intervention to encourage shared decision making.

Table 3. Mean Decision Process Scores from a SDP demonstration site, three "usual care" sites and a cross-section sample of Medicare patients for decision for how to treat breast cancer

Data source	N	Mean Process Score (SD)	t (comparing with demonstration site)	P
SDP Demonstration site	40	3.00 (.934)		
Usual care sites	227	2.54 (1.205)	2.7	<.01
Survey of Medicare beneficiaries treated for Br Ca	914	1.85 (1.25)	3.7	<.001

Table 3 shows that the SDP demonstration site patients reported a decision process that was much better than those clinical sites where there was no intervention to promote decision making. The comparable data from the survey of Medicare patients describing their decision making process for breast cancer treatment were much lower still.

We have similar data for decision making around hip and knee replacement.

Table 4. Mean Decision Process Scores from a SDP demonstration and three "usual care" sites and a cross-section sample of adults who made decisions for how to treat arthritis of the hip or knee.

Data source	N	Mean Decision Process Score (SD)	t (comparing with demonstration site)	P
SDP Demonstration site	178	2.96 (1.04)		
Usual care sites	204	2.6 (1.06)	3.3	<.001
TRENDS National survey of adults who made decisions about knee or hip replacement	268	2.70 (1.17)	2.5	<.02

As in Table 3, we see in Table 4 that the SDP demonstration sites had significantly better process scores from their patients than sites with no shared decision making initiative and was better than the national sample reported as well.

Finally, a small study at a clinical site in Stillwater, Minnesota collected data using the Decision Process Score questions from patients who discussed treatment for benign prostatic hyperplasia (BPH) with their urologists. They started collecting these data before introducing decision aids and continued to collect them after the use of decision aids that encouraged shared decision making became routine in the practice. Table 5 shows the results. While the Decision Process Score was pretty good before the use of decision aids, it was significantly better after they were introduced.

Table 5. Mean Decision Process Scores before and after the introduction of decision aids into process of treatment decisions for BPH.

When data collected	N	Mean Decision Process Score (SD)	t (comparing before and after data)	P
Before use of decision aids	47	3.02(.794)	3.12	<.01
After use of decision aids began	16	3.63 (.619)		

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
In summary, we have data that show clearly that decision making on average in the US as measured by this score is not very good and that clinical sites that commit to improved decision making attain average scores from their patients that are much higher than average. We think this is one of relatively few instances in which outcome measures based on patient reports are clearly linked to the way that clinical practices are trying to interact with patients.

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

The exclusions for these measures only apply to two of the decisions. For mastectomy, we exclude those who have had previous surgery for breast cancer because that may indicate a situation where there are not reasonable medical alternatives. We exclude prophylactic mastectomy because it is not treating cancer. For PCI, we exclude those who had a recent heart attack, because there is enough evidence of life extension that the decision may be seen as skewed toward the intervention. Those who have had previous coronary artery interventions may also be in complex medical situations. What we are trying to do is restrict measure to those with stable angina, which is a condition for which there clearly are alternatives to PCI and for which the paradigm of shared decision making clearly applies. Thus, the exclusions are specifically targeted to focus on those patients for whom shared decision making is clearly appropriate.

We did not test effects of exclusions on measures because they are about the clinical appropriateness of the measure.

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
There are two reasons we are not recommending any kind of risk adjustment. First, and perhaps most important, there is no ethical basis for saying the standards for engaging in shared decision making for these preference-sensitive surgical decisions should vary by patient characteristics. Second, as the data on disparities above shows, we have not found any systematic differences in average decision making scores based on age, gender, education or ethnicity. So, in our experience, adjustments would not have any meaningful effect on results.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

As noted in the analyses above, we simply used t tests to assess differences between mean Shared Decision Process scores from patient who made decisions in practices that had implemented procedures to promote shared decision making and patients who made decisions in usual care, either based on data from practices or from our national surveys.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The practices using decision aids and promoting shared decision making consistently had significantly better scores on this measure. The exception is decisions about surgery for lower back pain, which consistently get very high scores in "usual care". We think that is not a reflection of a problem with the measure but a reflection of the way back surgery decisions are made.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

We think the evidence is pretty strong that this measure validly reflects the extent to which a clinical practice is practicing shared decision making.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or*

eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)na

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
NA

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
In our experience, it is relatively rare for respondents to the follow-surveys not to answer all four questions. For example, from our test sites from patients making decisions about knee replacement and lower back surgery for herniated disc and spinal stenosis, the percentages of respondents not answering all four questions were 3%, 5% and 0% respectively; the percentages having more than one missing response were <1%, 2% and 0% respectively, from a total of 411 respondents. We did not experiment with alternative ways of handling missing data, because it really could not affect the results. We think leaving out anyone not answering all the questions or imputing a .5 score for one missing response (and eliminating anyone with more than one missing answer) would both be reasonable approaches to dealing with missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

See above

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

See 2b7.1 above

Bibliography

1. Bollen K, Lennox R. Conventional wisdom on measurement: A structural equation perspective. *Psychol Bull*. 1991;110(2):305-314. http://psycnet.apa.org/psycinfo/1992-03966-001.

2. Pass M, Belkora J, Moore D, Volz S, Sepucha K. Patient and observer ratings of physician shared decision making behaviors in breast cancer consultations. *Patient Educ Couns*. 2012;88(1):93-99. doi:10.1016/j.pec.2012.01.008.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Patient Reporting

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
It could be done online but response rates often are low with that approach. The rationale for using mail surveys with telephone reminder calls is to maximize the rate of return.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those

whose performance is being measured.
These questions have been extensively cognitively tested to ensure that they are consistently understood and that answers meaningfully describe patient experiences. We have used the questions proposed, and slight variations thereon, in a variety of survey designs: cross-section surveys of adults 40 and older, Medicare beneficiaries known to have had procedures based on claims, and clinical settings in which patients were identified by office staff or via medical records. The following observations have informed this proposal.
1. While it is preferable to ask about the degree to which providers discussed each reasonable option, the requirement to know the clinical details in order to ask about the reasonable alternatives makes that infeasible for many designs, plus, of course, the options have to change for each type of decision. The current question 4 can be used for basically any decision with minimal or no tailoring.
2. We can identify patients making decisions by asking them whether or not they had discussed an intervention, test or treatment. However, for cross-sections of adults or patients, the rates of any particular decision being made are too low to produce reliable data without very large samples.
3. We have tried to identify patients in clinical settings who faced a particular type of decision, regardless of which choice they ended up making. That would be the preferred way to measure decision quality at a clinical site. However, we have found it is very difficult to reliably identify patients who discuss an intervention but choose not to have it. Often such decisions do not end up in medical records, and there is no reasonable way to know which patients to ask about a decision. Physicians, who also would know that, have proven very unreliable at flagging patients who were in decision windows. Hence, in order to get comparable results across clinicians or clinical sites, we think the best option is to sample those patients who have actually had the target intervention. In that way, we think that patients can be reliably identified and we think the case is particularly strong that those who actually have an intervention should have had a good decision making process.
4. Because results are strongly related to which decision is being made, we think meaningful comparisons across sites can only occur for a particular decision.These questions have been extensively cognitively tested to ensure that they are consistently understood and that answers meaningfully describe patient experiences. We have used the questions proposed, and slight variations thereon, in a variety of survey designs: cross-section surveys of adults 40 and older, Medicare beneficiaries known to have had procedures based on claims, and clinical settings in which patients were identified by office staff or via medical records. The following observations have informed this proposal.
1. While it is preferable to ask about the degree to which providers discussed each reasonable option, the requirement to know the clinical details in order to ask about the reasonable alternatives makes that infeasible for many designs, plus, of course, the options have to change for each type of decision. The current question 4 can be used for basically any decision with minimal or no tailoring.
2. We can identify patients making decisions by asking them whether or not they had discussed an intervention, test or treatment. However, for cross-sections of adults or patients, the rates of any particular decision being made are too low to produce reliable data without very large samples.
3. We have tried to identify patients in clinical settings who faced a particular type of decision, regardless of which choice they ended up making. That would be the preferred way to measure decision quality at a clinical site. However, we have found it is very difficult to reliably identify patients who discuss an intervention but choose not to have it. Often such decisions do not end up in medical records, and there is no reasonable way to know which patients to ask about a decision. Physicians, who also would know that, have proven very unreliable at flagging patients who were in decision windows. Hence, in order to get comparable results across clinicians or clinical sites, we think the best option is to sample those patients who have actually had the target intervention. In that way, we think that patients can be reliably identified and we think the case is particularly strong that those who actually have an intervention should have had a good decision making process.
4. Because results are strongly related to which decision is being made, we think meaningful comparisons across sites can only occur for a particular decision.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
None

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance

results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Professional Certification or Recognition Program BCBS of MA Blue Distinction Specialty Care Program, The Alliance Quality Path recognition program (for hip and knee replacement)* http://www.the-alliance.org/uploadedFiles/Providers/QualityPath_knee_and_hip_replacement_RFP.pdf Quality Improvement (Internal to the specific organization) Partners Healthcare http://www.massgeneral.org/decisionsciences/assets/pdfs/OAKnee_DQI_SV.pdf Partners Healthcare http://www.massgeneral.org/decisionsciences/assets/pdfs/OAHip_DQI_SV.pdf

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Quality Path Program sponsored by the Alliance specifies measurement of shared decision making as part of their criteria for recognition. The Alliance is a cooperative of employers that includes more than 240 members who provide self-funded health benefits to more than 100,000 individuals. The network lets members choose from more than 80 hospitals, 13,500 total professional service providers, and 3,400 medical clinic sites in Wisconsin, Illinois, and Iowa. The purpose of the Quality Path program is to recognize providers and hospitals who are delivering high quality surgical care. The relevant section from the program detailing use of the measure is excerpted below and the entire program details can be found at:
See for example http://www.the-alliance.org/uploadedFiles/Providers/QualityPath_knee_and_hip_replacement_RFP.pdf

16 Decision Quality Assessment (p 18)
Supporting Documentation:
• Provide a description of the process for assessing the quality of shared decision making. This process needs to use the decision quality assessment tool available at:
o Knee: http://www.massgeneral.org/decisionsciences/assets/pdfs/OAKnee_DQI_SV.pdf
o Hip: http://www.massgeneral.org/decisionsciences/assets/pdfs/OAHip_DQI_SV.pdf

BlueCross BlueShield of Massachusetts Blue Distinction Specialty Care Program. Lists “Shared Decision Making” as one of their spine and hip/knee outcomes and specifies that “the program employs SDM processes and solicits patient feedback about their SDM process and uses a formal tool to conduct SDM.”

Partners Healthcare Population Health Management. The PHM group has developed a provider order entry tool with the goal of increasing accountability for surgical procedures. The tool is focused on elective procedures (PCI, hip and knee replacement, spine surgery) and requires clinicians document clinically appropriateness criteria and use of shared decision making.

These various initiatives have not been using this proposed process measure, but they are examples of places that are concerned about evaluating the quality of decision making and a certified measure of the Shared Decision Making Process would be the

obvious measure for them to use in their evaluation programs.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
New measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The Foundation is not in position to sponsor or implement quality accountability measurement. However, we think the efforts described in 4.1 are examples of the kinds of programs that will want to start using this measure when it is NQF approved.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
NO

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

1741 : Patient Experience with Surgical Care Based on the Consumer Assessment of Healthcare Providers and Systems (CAHPS)® Surgical Care Survey

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The approved PCMH and ACO CAHPS measures of shared decision making were adaptations of the measures we developed and are proposing. Those measures were used for respondents who reported they had discussed starting or stopping a prescription medication (for PCMH) and for patients who reported discussion a prescription medication or a procedure with a provider (ACO). The problem with integrating this measure into the CAHPS protocols includes both sample sizes and sample designs. This measure works best when applied to a specific kind of decision (eg. Decision to take medication for high blood pressure or decision to have surgery for herniated disc.) CAHPS samples relatively small numbers of ambulatory patients from a clinician's practice or a clinical site. Those samples do not include enough encounters at which decisions are made about specific medications or specific tests or surgical procedures to provide reliable data. Hence, they had to ask about any decisions about starting or stopping medications or surgical procedures and combine the answers for each type of decision. The numbers of such decisions tend to be very small, even when all medications or procedures are combined. Moreover, we have abundant data showing that the Shared Decision Making Process Score varies widely from medication to medication and procedure to procedure. (Zikmund=Fisher et al, 2010; Fowler et al, 2012; Fowler et al, 2014). The approach we are proposing, sampling patients who have undergone a procedure, provides the ability to control the sample sizes of respondents and provides for collecting data about the same decision when using the data to compare clinical sites—which is essential in order to meaningfully interpret the results as measures of quality of care.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
There is no other measure that we have identified of the shared decision process that has NQF endorsement. There was a shared decision making measure for back pain that consisted of whether or not physicians recorded in the medical record that they had reviewed various aspects of risks and benefits of back surgery prior to surgery. This measure is no longer endorsed. In addition, obviously patient reports of their discussions with physicians are very different from physician reports of their own perceptions of their discussions. We certainly think that patient reports are a more credible measure of what transpired.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQF_Measure_Steward_Agreement_IMDF_signed.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Informed Medical Decisions Foundation, a division of Healthwise
Co.2 **Point of Contact:** Floyd, Fowler, fjfowler@healthwise.org, 617-367-2000-
Co.3 **Measure Developer if different from Measure Steward:**
Co.4 **Point of Contact:** Floyd, Fowler, fjfowler@healthwise.org, 617-367-2000-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This was not the product of a formal work group. It emerged from an ongoing effort since 2007 by the Informed Medical Decisions Foundation to develop measures of decision quality. The following researchers played a significant role at one or more points in the development process.

Brian Zikmund-Fisher, Angie Fagerlin, Mick Couper and Eleanor Singer, all at the Survey Research Center at the University of Michigan worked with the Foundation staff on the first versions of the questions to measure the decision making process. Results were published in a number of papers from the DECISIONS survey, a national survey of adults 40 or older who had made decisions.
Karen Sepucha, at Massachusetts General Hospital, has been a central part of the research team in this area from the beginning.
Carol Cosenza at the Center for Survey Research at UMass Boston has worked on cognitive testing of these questions in various forms.

Floyd J Fowler, Jr and Carrie Levin at the IMDF have played the lead roles in coordinating this work.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?** As needed
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:** Citations:
1. Stacey D, Légaré F, Col NF, et al. Decision aids for people facing health treatment or screening decisions. Cochrane database Syst Rev. 2014;1:CD001431. http://www.ncbi.nlm.nih.gov/pubmed/24470076. Accessed December 14, 2014.
2. Fowler FJ, Gerstein BS, Barry MJ. How patient centered are medical decisions?: Results of a national survey. JAMA Intern Med. 2013;173(13):1215-1221. doi:10.1001/jamainternmed.2013.6172.
3. Fagerlin A, Sepucha KR, Couper MP, Levin CA, Singer E, Zikmund-Fisher BJ. Patients' knowledge about 9 common health conditions: the DECISIONS survey. Med Decis Making. 30(5 Suppl):35S - 52S. doi:10.1177/0272989X10378700.
4. Wexler RM, Gerstein BS, Brackett C, Fagnan LJ, Fairfield KM, Frosch DL LC, , Morrissey L, Simmons LH SD, Chang Y FF and BM. Patient Responses to Decision Aids in the United States. Int J Pers Cent Med. 2015;5(3):105-111.
5. Zikmund-Fisher BJ, Couper MP, Singer E, et al. Deficits and variations in patients' experience with making 9 common medical decisions: the DECISIONS survey. Med Decis Making. 30(5 Suppl):85S - 95S. doi:10.1177/0272989X10380466.
6. Fowler FJ, Gallagher PM, Bynum JPW, Barry MJ, Lucas FL, Skinner JS. Decision-making process reported by Medicare patients who had coronary artery stenting or surgery for prostate cancer. J Gen Intern Med. 2012;27(8):911-916. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3403150&tool=pmcentrez&rendertype=abstract. Accessed June 1, 2015.
7. Fowler FJ, Gerstein BS, Barry MJ. How patient centered are medical decisions?: Results of a national survey. JAMA Intern Med. 2013;173(13):1215-1221. http://www.ncbi.nlm.nih.gov/pubmed/23712194. Accessed April 9, 2015.

1b.3. <u>If no or limited performance data on the measure as specified is reported in 1b2</u>, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance **on the specific focus of measurement.** Include citations.

The best data on the opportunity and need for improvement come from national surveys of patients who either were known to have had procedures, based on Medical claims, or said they had made decisions with their doctors. For Medicare patients who had had prostate surgery for cancer or a PCI for coronary artery disease, the means scores on the Shared Decision Process Score were 2.7 and 1.2 respectively (out of a possible score of 4). (Fowler et al, 2012). In a national survey of adults 40 and older who reported having made one of more decisions about cancer screening, taking prescription medications and surgery, the mean SDP scores ranged from 1.5 (for mammograms) to 3.2 for back surgery. Only back surgery decision making (mean = 3.2) met our standard for a satisfactory decision making process (Fowler, Gerstein and Barry, 2013).

The following table summarizes means and SDs for scores from national samples:
Table 1: Mean Shared Decision Making Process Scores for 7 Common Surgical Procedures

Procedure	Mean Shared Decision Making Process Score	Standard Deviation	Data Source
Prostatectomy for Prostate Cancer	2.7	1.0	Survey of Medicare Patients who had surgery
Mastectomy for Breast Cancer	1.9	1.3	Survey of Medicare patients who had had surgery
PCI for coronary artery disease	1.2	1.0	Survey of Medicare patients who had had surgery
Hip replacement for osteoarthritis of the hip	2.5	1.2	National survey of adults 40 or older from Knowledge Networks panel (TRENDS)

Knee replacement for osteoarthritis of the knee	**2.8**	**1.1**	**National survey of adults 40 or older from Knowledge Networks panel (TRENDS)**
Surgery for lower back pain (disco or stenosis)	**3.2**	**1.0**	**National survey of adults 40 or older from Knowledge Networks panel (TRENDS)**

The evidence for the value of clinical practices devoted to shared decision making and that the SDP score is a valid measure of clinical performance comes from a number of studies of decision making in clinical practices, some of which were trying to implement shared decision making on a routine basis and using decision aids for many decisions. The following summarizes those results.

We have compared the aggregate Process Scores from patients treated a clinical sites that have committed to shared decision making, usually by including the routine use of decision aids, with reports of national cross-sections of patients from the TRENDS survey who made the same decisions.
Table 2. Mean Decision Process Scores at SDP Demonstration sites and from a national sample of patients for three orthopedic procedures.

Data Source	Decision Topic	N	Mean Process Score	Std. Deviation
TRENDS	Surgery: Knee Pain	163	2.81	1.139
Demo Sites	Knee Osteoarthritis	239	3.24**	.840
TRENDS	Surgery: Hip Pain	57	2.45	1.236
Demo Sites	Hip Osteoarthritis	129	3.31***	.864
TRENDS	Surgery: Low Back Pain	152	3.23	1.016
Demo Sites	Herniated Disc + Spinal Stenosis	55	3.38	.828

**p < .01
*** P< .001

For osteoarthritis of the knee and hip, it can be seen that the patients in practices where decision aids are used reported significantly better decision processes than a cross-section sample of adults who faced the same decisions. The responses did not differ for conversations about lower back pain, but the decisions about back pain were by far the best decision processes based on respondent reports in the national survey.

Because the data in the above table were collected with quite different time periods between the decision and the measurement, a better test may come from studies of breast cancer decision making in four clinical sites. One of these four sites routinely used decision aids and had support for patients when they met with their surgeons to facilitate getting patients' questions asked and answered. The other three sites practiced usual care, with no special intervention to encourage shared decision making.

Table 3. Mean Decision Process Scores from a SDP demonstration site, three "usual care" sites and a cross-section sample of Medicare patients for decision for how to treat breast cancer

Data source	N	Mean Process Score (SD)	t (comparing with demonstration site)	P
SDP Demonstration site	40	3.00 (.934)		
Usual care sites	227	2.54 (1.205)	2.7	<.01
Survey of Medicare beneficiaries treated for Br Ca	914	1.85 (1.25)	3.7	<.001

Table 3 shows that the SDP demonstration site patients reported a decision process that was much better than those clinical sites where there was no intervention to promote decision making. The comparable data from the survey of Medicare patients describing their decision making process for breast cancer treatment were much lower still.

We have similar data for decision making around hip and knee replacement.

Table 4. Mean Decision Process Scores from a SDP demonstration and three "usual care" sites and a cross-section sample of adults who made decisions for how to treat arthritis of the hip or knee.

Data source	N	Mean Decision Process Score (SD)	t (comparing with demonstration site)	P
SDP Demonstration site	178	2.96 (1.04)		
Usual care sites	204	2.6 (1.06)	3.3	<.001
TRENDS National survey of adults who made decisions about knee or hip replacement	268	2.70 (1.17)	2.5	<.02

As in Table 3, we see in Table 4 that the SDP demonstration sites had significantly better process scores from their patients than sites with no shared decision making initiative and was better than the national sample reported as well.

Finally, a small study at a clinical site in Stillwater, Minnesota collected data using the Decision Process Score questions from patients who discussed treatment for benign prostatic hyperplasia (BPH) with their urologists. They started collecting these data before introducing decision aids and continued to collect them after the use of decision aids that encouraged shared decision making became routine in the practice. Table 5 shows the results. While the Decision Process Score was pretty good before the use of decision aids, it was significantly better after they were introduced.

Table 5. Mean Decision Process Scores before and after the introduction of decision aids into process of treatment decisions for BPH.

When data collected	N	Mean Decision Process Score (SD)	T (comparing before and after data)	P
Before use of decision aids	47	3.02(.794)	3.12	<.01
After use of decision aids began	16	3.63 (.619)		

In summary, we have data that show clearly that decision making on average in the US as measured by this score is not very good and that clinical sites that commit to improved decision making attain average scores from their patients that are much higher than average.

1b.4. Provide disparities data from the measure as specified (**current and over time**) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
We have data that enable us to address the disparities issue for 6 of the 7 decisions. However, we have to combine herniated disc and stenosis surgery together into surgery for lower back pain. We do not have adequate numbers to do these analyses for hip replacement surgery. The data come from surveys of Medicare beneficiary surgical patients in 2008 (Prostate surgery, mastectomy and PCI) and from the TRENDS survey of a general population who said they had discussed or had knee replacement surgery or surgery for lower back pain, conducted in 2012. The covariates we looked at were age, education, race/ethnicity and gender. The following 5 tables show the results.

Table 6: PROCEDURE: Mastectomy

VARIABLE	GROUP	MEAN SHARED DECISION PROCESS SCORE	P	N
EDUCATION	COLLEGE GRAD	2.13	.005	63
	NOT COLLEGE GRAD	1.61		303
RACE	NON-HISPANIC WHITE	1.74	.068	328
	OTHER RACES	1.67		52
AGE	<65	NA		
	65+			
GENDER	MALE			
	FEMALE			

Table 7: PROCEDURE: SURGERY FOR PCA

VARIABLE	GROUP	MEAN SHARED DECISION PROCESS SCORE	P	N
EDUCATION	COLLEGE GRAD	2.81	.001	262
	NOT COLLEGE GRAD	2.53		398
RACE	NON-HISPANIC WHITE	2.68	.001	608
	OTHER RACES	2.24		65
AGE	<65	NA		
	65+			
GENDER	MALE			
	FEMALE			

Table 8: PROCEDURE: PCI (Stents)

VARIABLE	GROUP	MEAN SHARED DECISION PROCESS SCORE	P	N
EDUCATION	COLLEGE GRAD	1.29	.378	112
	NOT COLLEGE GRAD	1.21		408
RACE	NON-HISPANIC WHITE	1.21	.106	468
	OTHER RACES	1.43		51
AGE	<65	NA		
	65+			
GENDER	MALE	1.26	.119	348
	FEMALE	1.13		185

Table 9: PROCEDURE: Knee Replacement Surgery

VARIABLE	GROUP	MEAN SHARED DECISION PROCESS SCORE	P	N
EDUCATION	COLLEGE GRAD	3.03	.155	41
	NOT COLLEGE GRAD	2.74		122

RACE	NON-HISPANIC WHITE	2.74	.074	128
	OTHER RACES	3.06		35
AGE	<65	2.74	.403	85
	65+	2.89		78
GENDER	MALE	3.01	.026	81
	FEMALE	2.61		83

Table 10: PROCEDURE: Back surgery

VARIABLE	GROUP	MEAN SHARED DECISION PROCESS SCORE	P	N
EDUCATION	COLLEGE GRAD	3.27	.839	26
	NOT COLLEGE GRAD	3.23		126
RACE	NON-HISPANIC WHITE	3.07	.000	116
	OTHER RACES	3.77		36
AGE	<65	3.12	.055	93
	65+	3.42		58
GENDER	MALE	3.32	.310	76
	FEMALE	3.15		76

For several of the comparisons, the number of cases for one of the groups is less than 50, which limits the power to detect significant differences. In 2 of the 5 tables, those who completed college reported significantly higher Shared Decision Process scores than those with less formal education. All of the tables show race-related differences at or near the level needed for statistical significance. However, the differences go in different directions, with the non-Hispanic whites reporting better processes in some cases, worse in others. The Medicare surveys did not permit comparisons by age, and two of the procedures are gender-specific. Of the comparisons we could do with these groups, the males reported a better decision process than females for knee replacement, and those over 65 reported a better decision process than younger patients for deciding on back surgery. Although there is consistent evidence that patient levels of formal education are related to measures of patient knowledge (e.g Fagerlin et al, 2010) we do not have much evidence that racial or education groups consistently differ in the their reported interactions with providers about surgical decisions.

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare). List citations in 1c.4.

The table below, derived from data from the Dartmouth Atlas, shows the rates at which these procedures are performed and the variations in the rates at which they are performed for the Medicare population in 2012. The large differences between the high and low rate areas, exceeding a ten to one ratio in several cases, is widely interpreted as evidence that decisions are being driven by providers, not patients, and reflecting highly different physician ideas about how aggressively to use the procedures. Thus, in addition to the large number of procedures involved, this is compelling evidence of a need for greater patient involvement in decision making for these procedures. It should be noted that the numbers below understate the total procedures done, as they do not include procedures for those under 65. Prostate cancer surgery, mastectomy and surgery for herniated disc are particularly common among those under 65.

Table 11: RATES AND VARIATIONS OF SEVEN SURGICAL PROCEDURES FOR MEDICARE BENEFICIARIES IN (MOST RECENT YEAR 2012)

PROCEDURE	NATIONAL NUMBER OF PROCEDURES (2012)	NATIONAL RATES/1000 (2012)	RATE/1000 HIGH AREA (2012)	RATE/1000 LOW AREA (2012)
TOTAL KNEE REPLACEMENT	420,197	8.5	18.6	2.2
TOTAL HIP REPLACEMENT	197,740	4.0	7.5	0.6
BACK SURGERY (COMBINES STENOSIS AND DISC)	232,344	4.7	13.4	1.3
RADICAL PROSTATECOMY FOR PROSTATE CANCER	24,470	1.1	2.13	0.32
MASTECTOMY FOR BREAST CANCER	15,769	.58	1.01	0.20
PCI FOR CORONARY ARTERY DISEASE OR STABLE ANGINA	307,485	6.22	23.1	1.8

Citations

1. Stacey D, Légaré F, Col NF, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane database Syst Rev*. 2014;1:CD001431. http://www.ncbi.nlm.nih.gov/pubmed/24470076. Accessed December 14, 2014.
2. Fowler FJ, Gerstein BS, Barry MJ. How patient centered are medical decisions?: Results of a national survey. *JAMA Intern Med*. 2013;173(13):1215-1221. doi:10.1001/jamainternmed.2013.6172.

3. Fagerlin A, Sepucha KR, Couper MP, Levin CA, Singer E, Zikmund-Fisher BJ. Patients' knowledge about 9 common health conditions: the DECISIONS survey. *Med Decis Making*. 30(5 Suppl):35S - 52S. doi:10.1177/0272989X10378700.
4. Wexler RM, Gerstein BS, Brackett C, Fagnan LJ, Fairfield KM, Frosch DL LC, , Morrissey L, Simmons LH SD, Chang Y FF and BM. Patient Responses to Decision Aids in the United States. *Int J Pers Cent Med*. 2015;5(3):105-111.
5. Zikmund-Fisher BJ, Couper MP, Singer E, et al. Deficits and variations in patients' experience with making 9 common medical decisions: the DECISIONS survey. *Med Decis Making*. 30(5 Suppl):85S - 95S. doi:10.1177/0272989X10380466.
6. Fowler FJ, Gallagher PM, Bynum JPW, Barry MJ, Lucas FL, Skinner JS. Decision-making process reported by Medicare patients who had coronary artery stenting or surgery for prostate cancer. *J Gen Intern Med*. 2012;27(8):911-916. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3403150&tool=pmcentrez&rendertype=abstract. Accessed June 1, 2015.
7. Fowler FJ, Gerstein BS, Barry MJ. How patient centered are medical decisions?: Results of a national survey. *JAMA Intern Med*. 2013;173(13):1215-1221. http://www.ncbi.nlm.nih.gov/pubmed/23712194. Accessed April 9, 2015.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2967
Measure Title: Home and Community Based Services (HCBS) Experience of Care (EoC) Measures
Measure Steward: Centers for Medicare and Medicaid Services
Brief Description of Measure: Home and Community Based Services (HCBS) Experience of Care (EoC) measures derive from a cross disability survey to elicit feedback from adult Medicaid beneficiaries receiving home and community based services (HCBS) about the quality of the long-term services and supports they receive in the community.

The measures consist of seven scale measures, 6 global rating and recommendation measures and 6 individual measures:
Scale Measures
1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Ratings Measures
8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendations Measures
11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends –– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)
Developer Rationale: Scale Measures
Staff are Reliable and Helpful. Assessing the performance of Medicaid direct care providers (i.e., personal assistants, behavioral health staff, homemakers) from the perspective of the beneficiary is important in evaluating the quality of services they render. This measure is based on beneficiary assessment of direct care staff reliability (showing up on time, stay as long as supposed to, communicate absences) and sensitivity to their privacy needs during the provision of personal care.

Staff Listen and Communicate Well. This measure is based on beneficiary assessment of direct care staff's communication skills and responsiveness to the person's needs. Specifically communication in a way that is understood by the beneficiary, respectful, and staff who listen carefully to what the beneficiary needs/wants and who, therefore, understand what the beneficiary needs. This is essential to the delivery of person-centered care and support. Person-centered care and support is required in Medicaid HCBS programs (Federal Register: https://federalregister.gov/a/2014-00487).

Case Manager Is Helpful. In HCBS programs, the case manager is responsible for monitoring the beneficiary's receipt of services and supports to ensure the service plan is being implemented as specified and that the person's needs are being adequately met. In order to meet these requirements, the case manager must be available to the beneficiary when s/he contacts him/her, and responsive to their changing/emerging needs. This measure is based on the beneficiary's assessment of case manager accessibility and responsiveness.

Choosing Services That Matter to You. A basic tenet of Medicaid HCBS services is that the beneficiary is involved in choosing their services/supports so that the service plan is truly person-centered, and that direct care staff implement the service plan in a person-centered manner. This measure is based on the beneficiary's assessment of the extent to which their service plan and direct care workers are person-centered.

Transportation to Medical Appointments. The health and welfare of beneficiaries must be ensured in the delivery of Medicaid HCBS (42 CFR §441: 302). Integral to assuring the health of beneficiaries is getting to medical appointments. This composite is based on the beneficiary's assessment of the extent to which they have transportation to medical appointments, whether the transportation provider is reliable, and whether the transportation is sufficiently accessible.

Planning your time and activities. Medicaid home and community-based services and supports should facilitate outcomes that are consistent with allowing beneficiaries to live the lives they choose – both in terms of daily routine as well as socializing with family and friends, and engaging in community activities. This measure is based on the beneficiary's assessment of the extent to which they have choice and control over these aspects of their lives.

Personal Safety and Respect. Beneficiaries of Medicaid HCBS should be assured that HCBS providers treat them with respect, that they will not be financially exploited by providers coming into their homes, and that they have someone to go to if they are treated badly. This measure will help HCBS programs assess this aspect of program quality.

Individual Item Measures
Global Ratings of Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) – separate measures per staff type. In concert with more specific measures and scale measures, global ratings provide additional information for assessing program quality and can be used as a metric in evaluating quality improvement.

Would Recommend Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) to Family and Friends –separate measures per staff type. Beneficiaries' recommendation are yet another aspect of global experience with a program, and can be used for evaluating program quality and in quality improvement initiatives.

Individual Unmet Need Measures:

- Unmet Need in Dressing/bathing Due to Lack of Help
- Unmet Need in Meal Preparation/Eating Due to Lack of Help
- Unmet Need in Medication Administration Due to Lack of Help
- Unmet Need in Toileting Due to Lack of Help
- Unmet Need with Household Tasks Due to Lack of Help

None of the Unmet Need items were captured in a scale measure because they did not correlate with each other in factor analysis. But the advisory panel for the measures development strongly recommended all unmet need stand-alone items be treated as individual measures as the evaluation of unmet need in HCBS is critically important for determining program quality. One of the most basic reasons for the existence of HCBS programs is to meet self-care needs (bathing, dressing, toileting, medication administration) and needs that, if not met, make successful community living untenable (meal preparation/eating, cleaning/laundry). These measures are intended for use in assessing program quality and for quality improvement initiatives.

Hit or Hurt by Staff. This item was not retained in the Personal Safety and Respect scale measure due to low variation within responses. However, the advisory panel for the measures development felt this measure is important for establishing the personal safety of program beneficiaries, as physical abuse by staff is a "never event" that should be tracked in any HCBS quality management system.

Numerator Statement: HCBS service experience is measured in the following areas. Attached Excel Table S.2b includes the specific item wording for each measure and the response options that go into the numerator.

Scale Measures

1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Rating Measures

8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendation Measures

11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends –– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures

14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Denominator Statement: The denominator for all measures is the number of survey respondents. Individuals eligible for the HCBS survey include Medicaid beneficiaries who are at least 18 years of age in the sample period, and have received HCBS services for 3 months or longer. Eligibility is further determined using three cognitive screening items, administered during the interview:

Q1. Does someone come into your home to help you? (Yes, No)
Q2. How do they help you?
Q3. What do you call them?

Individuals who are unable to answer these cognitive screening items are excluded. Some measures also have topic-specific screening items as well. Additional detail is provided in S.9.

Denominator Exclusions: Individuals less than 18 years of age and individuals that have not received HCBS services for at least 3 months should be excluded. During survey administration, additional exclusions include individuals that failed any of the cognitive screening items mentioned in the denominator statement below.

Measure Type: PRO
Data Source: Patient Reported Data/Survey
Level of Analysis: Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

This submission contains information for 19 Patient Reported Outcome Performance Measures (PRO-PMs) derived from the Home and Community Based Services (HCBS) Experience of Care (EoC) survey. The measures consist of seven scale measures, 6 global rating and recommendation measures, and 6 individual measures:

Scale Measures
1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Ratings Measures
8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendations Measures
11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends –– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Summary of evidence:

- The developer provides a diagram that illustrates the path to potential beneficiary outcomes starting with the key processes (i.e., person-centered assessment and service planning) and resulting services (i.e., HCBS services and supports) that are expected to influence the beneficiary assessment of services/supports as well as beneficiary outcomes. Although not stated explicitly, these activities likely also would affect overall ratings of the care provided and willingness to recommend the HCBS services and supports.
- To assess if the target population values the measured PROs and find them useful, the developer utilized input from the HCBS beneficiary audience as well as stakeholders in the broader HCBS community. They state that the audiences have consistently supported the proposed measures as necessary and important.
- This input included focus groups and interviews, public comment via the Federal Register, and a Federal Advisory Panel.

Guidance from the Evidence Algorithm
Pro-based measure (Box 1) →Relationship between the outcome and at least one healthcare action is identified and supported by the rationale (Box 2) → PASS

Question for the Committee:
- Is there at least one thing that the provider can do to achieve a change in the measure results?
- Does the Committee agree that HCBS patients value queries about the various domains included in the HCBS Experience of Care survey?

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- See tables 1b.2a and 1b.2b in attached tables.

- Performance data were calculated through the testing of the measure and were provided at both the summary score (measure) and item (question) level. The data provided was collected from March – October, 2013 and consist of data from 26 Medicaid HCBS programs across 10 states. Performance data on the individual items used for the various measures are included in the supplementary materials.

Measures	Mean	Standard Deviation	25th Percentile	50th Percentile	75th Percentile
Staff are reliable and helpful	93.23	3.5	91.69	93.86	95.51
Staff listen and communicate well	93.06	2.44	91.5	93.06	94.87
Case manager is helpful	92.06	3.97	89.34	91.19	95.64
Choosing the services that matter to you	87.33	4.98	84.91	87.97	90.18
Transportation to medical appointments	90.49	4.36	87.21	90.35	93.59
Personal safety and respect	97.4	1.24	96.48	97.56	98.23
Planning your time and activities	81.75	2.42	80.92	81.89	83.48
Global Rating of Personal Assistance/Behavioral Health Staff	89.88	4.21	88.77	90.04	92.04
Global Rating of Homemaker	88.93	5.51	88.7	90.57	91.57
Global Rating of Case Manager	87.29	4.85	85.51	88.64	90.21
Recommendation of Personal Assistance/Behavioral Health Staff	88.84	5.65	87.89	88.9	91.72
Recommendation of Homemaker	86.59	9.6	85.05	89.14	93.29
Recommendation of Case Manager	86.3	4.37	84.07	86.81	88.74
Unmet need in dressing/bathing	34.19	24.93	17.15	35.85	52.45
Unmet need in meal preparation/eating	37.76	28.52	15.39	36.34	51.34
Unmet need in medication administration	72.41	27.1	62.99	77.42	94.45
Unmet need in toileting	95.82	5.51	94.23	96.81	100
Unmet need with household tasks	52.97	23.57	37.89	52.97	70.57
Physical Safety Measure: Hit or hurt by staff	99.69	0.67	99.81	99.96	100.00

Disparities

- The developer indicates the measures in the submission focus on people who are elderly with disabilities, individuals with physical disabilities, persons with intellectual/developmental disability, individuals with brain injury, and those with serious mental illness. who receive Medicaid-funded home and community-based services. As such, the target population mirrors those in a typical Medicaid population with evidence of disparities due to lower income, race and ethnicity.
- Tables 1b.4a, 1b.4b, 1b.4c, and 1b.4d provide summary statistics for the measure groupings for these populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement:
Scale Measures

1. Staff are reliable and helpful ☐ **High** ☐ **Moderate** X **Low** ☐ **Insufficient**
2. Staff listen and communicate well ☐ **High** ☐ **Moderate** X **Low** ☐ **Insufficient**
3. Case manager is helpful ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**
4. Choosing the services that matter to you ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**
5. Transportation to medical appointments ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**
6. Personal safety and respect ☐ **High** ☐ **Moderate** X **Low** ☐ **Insufficient**

7. Planning your time and activities ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

Global Ratings Measures

8. Global rating of personal assistance and behavioral health staff ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

9. Global rating of homemaker ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

10. Global rating of case manager ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

Recommendations Measures

11. Would recommend personal assistance/behavioral health staff to family and friends ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

12. Would recommend homemaker to family and friends ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

13. Would recommend case manager to family and friends ☐ **High** X **Moderate** ☐ **Low** ☐ **Insufficient**

Unmet Needs Measures

14. Unmet need in dressing/bathing due to lack of help X **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

15. Unmet need in meal preparation/eating due to lack of help X **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

16. Unmet need in medication administration due to lack of help X **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

17. Unmet need in toileting due to lack of help ☐ **High** ☐ **Moderate** X **Low** ☐ **Insufficient**

18. Unmet need with household tasks due to lack of help X **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Physical Safety Measure

19. Hit or hurt by staff ☐ **High** ☐ **Moderate** X **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**I would rate as moderately important. If I understand the specifications correctly, this measure is intended to assess an HCB "program", which I believe the developer means to be an entity, probably a company like the VNA, providing HCB services to the state's Medicaid population. The assessment of the quality of care being provided from the perspective of the patient could be valuable to the state in monitoring the quality of the service.

**Measurement at the global, scale and individual level to demonstrate qualitative and quantitative evidence related directly and tangentially to HCBS. Many of these quality and safety measures are required as part of a HCBS under Medicaid regulations-For example, transportation falls under federal regulation for Medicaid beneficiaries as transport to medical apts is part of their covered benefit. Measuring the quality if the service and availability of service is a tangential outcome of access and utilization of the service.

For Scale Measure: Is it possible to have a metric that shows patient activation and/or engagement in HCBS? For example, how empowered to patients feel that they can participate in and/or codesign their home care plan
Scale measure: Is there a way to add cultural competency and/or language? Ease of interpretation on interpreter services

1b. Performance Gap

Comments:

**Here I would rate the measure as barely moderate. Except for the unmet needs measures the performance gap is pretty narrow. I have some concerns about the appropriateness of the unmet needs measure since, generally, the volume and kind of HCB services are usually dictated by a plan of care that is determined by the state, not the program. The program probably cannot increase either without state approval. We should explore this issue with the developer when we discuss this component of 2967.

**Yes. Disparities due to lower income, race and ethnicity (the target population focused on Medicaid, people who are elderly with disabilities and individuals with physical disabilite4s and server mental illness)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Self-reports of Medicaid beneficiaries of home and community based services

Specifications:

- The measures is specified for the program level of analysis for home and community based services; higher scores are an indicator of better quality
- The measures that comprise this submission include scale measures (7), global ratings (3), recommendation ratings (3), unmet needs (5), and a physical safety measure (1). The attached spreadsheet contains the individual survey items and item mapping for each measure grouping
- The frequency of data collection/aggregation is at the discretion of state users. The developer notes that CMS has determined the survey from which the measures are derived will be conducted on a voluntary basis by states. It is anticipated that states would field the survey no more frequently than annually per HCBS program.
- The denominator is Medicaid beneficiaries who are at least 18 years of age in the sample period, and have received HCBS services for 3 months or longer.
- Eligibility is further determined using three cognitive screening items, administered during the interview (Individuals who are unable to answer these cognitive screening items are excluded):
  Q1. Does someone come into your home to help you? (Yes, No)
  Q2. How do they help you?
  Q3. What do you call them?
- The proposed provider-related measures in this submission focus on the most common provider types for adults receiving Medicaid HCBS. These include personal assistance providers, behavioral health staff, homemakers and case managers.
- Case-mix adjustment is done via regression methodology or a covariance adjustment. Case-mix adjustment is used to adjust scores for various patient and survey mode characteristics.
- Scoring specifications for the measures follow the same general scoring approach as used by other CAHPS surveys that use the CAHPS analysis program. The measures are based on case-mix adjusted means that are transformed into a 0–100 metric.
- Sampling should be stratified by HCBS program within each state, in order to allow comparisons of measure results for each HCBS program to the state mean. The source of the sample frame is the state Medicaid agency or an entity delegated by the state Medicaid agency (e.g., state agency other than the Medicaid agency that operates the program, a MCO, a case management agency, state county, etc.).
- Results suggest that the effective sample size should be 400 people per stratum (with smaller programs including the census).
- Due to the impairments (i.e., cognitive, hearing) prevalent among individuals served by HCBS programs, stakeholders recommend that the survey be conducted through in-person interviews. Based on field test results, administering the survey by phone was found appropriate if a statistical adjustment for survey mode is made for mixed-mode administrations.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The developers conducted a pilot test and a field test of the survey with 26 Medicaid HCBS programs across ten states. The 10 states were geographically dispersed and included AZ, CO, CT, GA, KY, LA, MD, MN, NH, and TN; these states (with the exception of TN) were CMS Testing Experience and Functional Tools (TEFT) Demonstration grantees
- There were 2,336 completed HCBS EoC surveys from 26 Medicaid HCBS programs included in the analysis of the survey data. The testing was conducted from October 2013 – March 2015
- Reference Exhibit 1. States, Populations, Programs, Authorities, and Total Returned Surveys
- Data element reliability was assessed using Cronbach's Alpha values which assess internal consistency of the survey items used in the scale measures.
- HCBS program-level reliability was assessed by determining inter-unit reliability (IUR). Unit-level reliability indicates the extent to which the experiences of respondents within a unit (e.g., HCBS program) correlate with one another compared to the amount that reported experiences differ among units. The developers indicate that one of the primary purposes of these measures is to be able to detect difference among HCBS programs, and thus, this ratio is a good indicator of the extent to which the scale measures and other survey items accomplish this goal.

Results of reliability testing

- Tab 1.b.2a in the supplementary tables file for item-level IUR statistics for survey items used in the scale measures

Measures	IUR
Unmet need in dressing/bathing	0.16
Unmet need in meal preparation/eating	0.46
Unmet need in medication administration	0.63
Unmet need in toileting	-0.28
Unmet need with household tasks	0.15
Physical Safety Measure: Hit or hurt by staff	0.36

Exhibit 2. Cronbach's Alpha Values for Scale Measures

Staff are reliable and helpful	0.84
Staff listen and communicate well	0.84
Case manager is helpful	0.82
Choosing the services that matter to you	0.50
Transportation to medical appointments	0.70
Personal safety and respect	0.17
Planning your time and activities	0.55

Exhibit 3. HCBS Inter-unit reliability (IUR) Statistics

Staff are reliable and helpful	0.66
Staff listen and communicate well	0.70
Case manager is helpful	0.38
Choosing the services that matter to you	0.77
Transportation to medical appointments	0.68

Personal safety and respect	0.32
Planning your time and activities	0.44
Overall Rating of Personal Assistance/Behavioral Health Staff	0.43
Would Recommend Personal Assistance/Behavioral Health to Family and Friends	0.55
Overall Rating of Homemaker	0.42
Would Recommend Homemaker to Family and Friends	0.76
Overall Rating of Case Manager	0.57
Would Recommend Case Manager to Family and Friends	0.48

- For Cronbach's alpha, 0.70 or higher is a widely-accepted rule of thumb for a set of items to be considered a scale.
  - The Cronbach's Alpha scores range from 0.84 to 0.17, with three measures falling below the recommended 0.70 threshold. These were Planning your time and activities (0.55), Choosing the services that matter to you (0.50), and Personal safety and respect (0.17). While these values are below the recommended threshold, the developer indicated these measures were all deemed critical by the technical expert panel for assessing the quality of a HCBS program.
- If the IUR is higher, the ability of the item or scale measure to discriminate across programs is greater. Scales with reliability coefficients above 0.70 provide adequate precision for use in statistical analysis of unit-level comparisons. As the IUR gets smaller, a larger sample is needed in order to reliably discriminate across programs.
  - The IUR values at the program level for the scale measures, global measures and recommendation measures range from 0.77 to 0.32, with the majority of measures (10/13) falling below the 0.70 threshold. This indicates that these measures will need a larger sample size to effectively discriminate among programs.
  - The IUR values at the program level for the unmet needs and physical safety measures range from -0.28 - 0.63.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical testing conducted with measure as specified (Box 2) → Score-level testing conducted (Box 4) → Method of testing appropriate (Box 5) → Moderate certainty that the scores are reliable for 8 measures; lower certainty for 11 measures, although reliability will likely be higher if number of respondents is higher (than 200).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability:

Scale Measures

1. Staff are reliable and helpful ☐ High X Moderate ☐ Low ☐ Insufficient
2. Staff listen and communicate well ☐ High X Moderate ☐ Low ☐ Insufficient
3. Case manager is helpful ☐ High X Moderate ☐ Low ☐ Insufficient
4. Choosing the services that matter to you ☐ High X Moderate ☐ Low ☐ Insufficient
5. Transportation to medical appointments ☐ High X Moderate ☐ Low ☐ Insufficient
6. Personal safety and respect ☐ High ☐ Moderate X Low ☐ Insufficient
7. Planning your time and activities ☐ High ☐ Moderate X Low ☐ Insufficient

Global Ratings Measures

8. Global rating of personal assistance and behavioral health staff ☐ High X Moderate ☐ Low ☐ Insufficient
9. Global rating of homemaker ☐ High ☐ Moderate X Low ☐ Insufficient
10. Global rating of case manager ☐ High ☐ Moderate X Low ☐ Insufficient

Recommendations Measures

11. Would recommend personal assistance/behavioral health staff to family and friends ☐ High ☐ Moderate X Low ☐ Insufficient

12. Would recommend homemaker to family and friends ☐ High X Moderate ☐ Low ☐ Insufficient

13. Would recommend case manager to family and friends ☐ High ☐ Moderate X Low ☐ Insufficient

Unmet Needs Measures

14. Unmet need in dressing/bathing due to lack of help ☐ High ☐ Moderate X Low ☐ Insufficient

15. Unmet need in meal preparation/eating due to lack of help ☐ High ☐ Moderate X Low ☐ Insufficient

16. Unmet need in medication administration due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient

17. Unmet need in toileting due to lack of help ☐ High ☐ Moderate X Low ☐ Insufficient

18. Unmet need with household tasks due to lack of help ☐ High ☐ Moderate X Low ☐ Insufficient

Physical Safety Measure

19. Hit or hurt by staff ☐ High ☐ Moderate X Low ☐ Insufficient

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ Yes ☐ Somewhat ☐ No

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ Measure score ☐ Data element testing against a gold standard ☐ Both

Method of validity testing of the measure score:

- **☐ Face validity only**
- **☒ Empirical validity testing of the measure score**

Validity testing method:

- Criterion validity refers to the extent to which the HCBS scale measures agree with some criterion of the "true" value of the measure, and can be predictive or concurrent. The developers estimated correlation coefficients between each global rating measure and each scale measure.
- The developers examined correlations among the scale measures to determine if they measure different constructs. As these are all measures of beneficiary experience with HCB services, the factors are expected to be related; however, all inter-scale measure correlations should be below 0.80 to indicate that these 7 factors, while related, do not overlap to the point of being redundant.

Validity testing results:

- If the scale measures have good concurrent validity, then they should have a moderate to strong correlation ($r > 0.30$) with a conceptually related global rating measure.

Correlation of Scale Measures and Related Global Rating Measures

Measure	Correlation with Global Rating of Personal Assistance Staff
Staff are reliable and helpful	0.36*
Staff listen and communicate well	0.37*
Personal safety and respect	0.24*
Measure	**Correlation with Global Rating of Homemaker**
Staff are reliable and helpful	0.29*
Staff listen and communicate well	0.33*
Personal safety and respect	0.19*
Measure	**Correlation with Global Rating of Case Manager**
Case manager is helpful	0.38*
Choosing the services that matter to you	0.33*

*p <.001

- For most measures, the correlations between the scale measures and the related global rating measures were moderate, suggesting that the scale measures are valid measures of beneficiary experience with these providers. The correlation for Personal Safety and Respect was low; however, it should be noted that there was not much variance in the items for this measure.

Inter-Scale Correlations

Final Scale Measures	Staff are reliable and helpful	Staff are reliable and helpful	Case manager is helpful	Choosing the services that matter to you	Transportation to medical appointments	Personal safety and respect	Planning your time and activities
Staff are reliable and helpful	1.00	-	-	-	-	-	-
Staff listen and communicate well	0.49	1.00	-	-	-	-	-
Case manager is helpful	0.24	0.21	1.00	-	-	-	-
Choosing the services that matter to you	0.12	0.12	0.11	1.00	-	-	-
Transportation to medical appointments	0.32	0.35	0.27	0.07	1.00	-	-
Personal safety and respect	0.22	0.32	0.28	0.11	0.23	1.00	-
Planning your time and activities	0.26	0.23	0.19	0.08	0.32	0.27	1.00

*All correlations are statistically significant at p <.001

- The scale measures were somewhat correlated with each other as they are all measures of beneficiary experience. However, no values were above 0.80, suggesting that these scales are measuring unique concepts.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

N/A – there are no "true" exclusions to these measures

Questions for the Committee:

- *Do you have any reasons/evidence to believe there should be exclusions to these measures?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☒ **Yes** ☐ **No**

Risk adjustment summary [Risk adjustment summary

- The developers tested the beneficiary characteristics of age, health status (both general health and emotional/mental health), gender, and whether the respondent lived alone as case-mix adjusters. These characteristics typically have the strongest and most consistent associations with patient-reported problems in other CAHPS surveys. In addition, they tested several survey design characteristics – survey mode.
- The research team used stepwise regression to select a subset of the potential case-mix adjusters for further analysis. Stepwise regression analyses evaluated the strength of the relationship of each potential adjuster to ten global rating and scale measures in separate models in which each measure was regressed on all of the potential adjusters.
- The research team then estimated the heterogeneity factor, predictive power, explanatory power, and impact factor for each potential case-mix variable selected in the regression models.
- Variables that had an impact factor >1.0, and were eligible to be considered as case- mix adjusters, included general health rating, mental health rating, age, gender, whether respondent lives alone, survey administration mode, and response option.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer *used t-tests to compare the case-mix adjusted mean scores of each item, scale score, and global rating for each HCBS program within a state to the mean score of all programs combined within the state. A p-value of <0.05 was used to determine whether the scores were statistically significantly different from each other.*
- Exhibit 9 in the testing form shows counts of programs that were statistically significantly different above or below their state mean for each measure. The exhibit also reports the percentage of programs that were statistically significant in either direction from their state mean.
- The developer summarizes that the findings demonstrate that the measures produce results that adequately discriminate between service recipients' experience of care in their program compared to all programs within a state.

Question for the Committee:

Does this measure identify meaningful differences about quality?

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The developers conducted a nonresponse bias analysis to evaluate whether respondents and nonrespondents differed significantly.

Exhibit 10. Sample Frame Demographic Characteristics

Characteristics	Nonrespondents n=13,940	Respondents n=1,624	Total (Nonrespondents and Respondents Combined) N=15,564
HCBS Population*			

Aged (65+)	34.0	31.0	33.7
Disabled (<65)	36.4	41.8	36.9
ID/DD	19.0	11.3	18.2
TBI	4.2	6.3	4.4
SMI	6.4	9.6	6.8
Primary Language			
English	97.1	97.7	97.2
Spanish	2.0	1.9	2.0
Other	0.9	0.4	0.8
Metropolitan Statistical Area*			
Yes	74.3	76.5	74.5
No	25.7	23.5	25.5
Gender			
Male	41.9	43.0	42.0
Female	58.2	57.0	58.0
Assigned Survey Response			
Alternate	50.1	49.0	49.9
Standard CAHPS	50.0	51.1	50.1
Assigned Survey Mode			
In-person	80.6	79.2	80.4
Phone	19.4	20.8	19.6
State†*			
AZ	9.4	11.4	9.6
CO	17.7	15.0	17.4
GA	14.1	16.2	14.3
MD	19.2	7.1	18.0
MN	14.5	23.7	15.4
NH	25.2	26.6	25.3
Guardian*			
Yes	10.3	4.0	9.7

No	89.7	96.0	90.4

*Nonrespondents and respondents significantly differ by this characteristics at p <0.05

Guidance from the Validity Algorithm
Specifications consistent with evidence (Box 1) →Threats to validity assessed (Box 2) → Empirical testing conducted for the measure as specified (Box 3) → Testing at the score-level conducted (Box 6) →High certainly that the scores are valid indicators of quality

Preliminary rating for validity:
Scale Measures
1. Staff are reliable and helpful ☐ High X Moderate ☐ Low ☐ Insufficient
2. Staff listen and communicate well ☐ High X Moderate ☐ Low ☐ Insufficient
3. Case manager is helpful ☐ High X Moderate ☐ Low ☐ Insufficient
4. Choosing the services that matter to you ☐ High X Moderate ☐ Low ☐ Insufficient
5. Transportation to medical appointments ☐ High X Moderate ☐ Low ☐ Insufficient
6. Personal safety and respect ☐ High X Moderate ☐ Low ☐ Insufficient
7. Planning your time and activities ☐ High X Moderate ☐ Low ☐ Insufficient

Global Ratings Measures
8. Global rating of personal assistance and behavioral health staff ☐ High X Moderate ☐ Low ☐ Insufficient
9. Global rating of homemaker ☐ High X Moderate ☐ Low ☐ Insufficient
10. Global rating of case manager ☐ High X Moderate ☐ Low ☐ Insufficient

Recommendations Measures
11. Would recommend personal assistance/behavioral health staff to family and friends ☐ High X Moderate ☐ Low ☐ Insufficient
12. Would recommend homemaker to family and friends ☐ High X Moderate ☐ Low ☐ Insufficient
13. Would recommend case manager to family and friends ☐ High X Moderate ☐ Low ☐ Insufficient

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient
15. Unmet need in meal preparation/eating due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient
16. Unmet need in medication administration due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient
17. Unmet need in toileting due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient
18. Unmet need with household tasks due to lack of help ☐ High X Moderate ☐ Low ☐ Insufficient

Physical Safety Measure
19. Hit or hurt by staff ☐ High X Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 and 2b.1 Specifications:
Comments:
**Specification around the cognitive screening questions- if this is done at the hospital (where people are assessed often for home service or in the post acute care setting) how might you control for false positives on the cognitive impairment screening? Knowing, this is a higher risk when people are in acute care settings??

For the denominator of Medicaid beneficiaries who are 18 or older and have had HCBS services for three month or longer, does this county resumption of care or is it aggregate three months in a certain amount of time?

Case-mix adjustment is done via regression methodology or a covariance adjustment.- How do we ensure quality of coding for risk adjusted revenue or is this out of scope of this measure?

Concern for implementation: how do we spread and scale, especially of the recommendations are for in person ? How easy will it be to regularly implement this survey? What about considerations of Medicaid churn?

2a2. Reliability Testing

Comments:

**Developer reports the survey was used for 26 different programs. Total respondents were 2336, an average of less than 100 per program. Is this sufficient to determine reliability?

We also need an explanation of the recommendation that results should be "stratified" in order to compare a program's score with the state mean. Stratified how: size of program? composition of caseload?

**Both measure score and data element 2336 completed surveys across 10 Medicaid SCHS service sites

Data element reliability was assessed using Cronbach's Alpha values. There needs to be a larger sample size to effectively discriminate among programs because the scale measures, global measures and recommendation measures range from .77-.32

2b2. Validity–Testing

Comments:

**For most measures, the correlations between the scale measures and the related global rating measures were moderate, suggesting that the scale measures are valid measures of beneficiary experience with these providers. The correlation for Personal Safety and Respect was low; however, it should be noted that there was not much variance in the items for this measure.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- It is recommended that the HCBS EoC Survey be administered in-person or by phone. CATI or CAPI data collection is recommended which allow for the creation of electronic databases post data collection.
- The developers include notes on opportunities to improve survey data collection learned from the field-test and recommendations on sampling and seasonality timing.
- The final HCBS EoC survey will be available to state Medicaid Agencies for use free of charge. In addition to the survey instrument, users will have access to comprehensive materials supporting fielding, analysis, and reporting as well as CAHPS Analysis Program that performs analysis and significance testing.

Questions for the Committee:

- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3 Feasibility

Comments:

**I am very doubtful that many states will want to use this measure. The data source is the responses from a 95 question survey which the developer recommends be administered in person, or possibly by phone. The suggested sample size is 400. Administering a survey of that length even by phone is expensive; it is unlikely that states will require

the program to do and pay for (HCB programs are generally not well funded) nor that the state will be able to pay for. I believe the states in the test received federal grants to cover the cost.

**My concern is spread and scale- how do large systems do interviews in person?. I am also concerned with interview responses creating bias. What is the plan for Medicaid churn and re-surveying patients? What about training and oversight of contracted agencies of whom there may be little power or influence to improve performance? Would this be a way to vet these agencies?

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
The measure is new and not currently in use, but public reporting and quality improvement uses are planned.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details The HCBS EOC survey is new, and so are the measures described in this submission. The survey is under review by the CAHPS Consortium for evaluation of use of the CAHPS trademark. Upon receipt, it is anticipated this survey and measures will be put into voluntary use by state programs for QI initiatives and service planning.

Improvement results New Measure

Unexpected findings (positive or negative) during implementation None identified

Potential harms None identified

Feedback:
Due to the newness of the measures in this submission and the recently completion of survey testing and analysis, the submission has not been viewed by other NQF bodies. However, both the MAP Duals Workgroup and the Home and Community Based Services Committee have been following the development and have expressed interest in the measurement set. They cite a paucity of measures for the HCBS care setting and the broader targeted populations that comprise this denominator.

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4 Usability and Use
Comments:

**I believe the scores are not meant to be publicly reported. The developer talks of their use to "compare with state mean", which suggest intended uses are for QI improvement on the part of the program and for QI oversight by the state. From consumer perspective, it would desirable for the results to be public in order to guide consumer selection of HCB program (if there is a choice in her region).

We should have a separate discussion of the physical safety measure: this veers close to the "never event" category and public reporting is a sensitive issue. We might wish to recommend some cautionary language if we decide to recommend endorsement.

**Would the public reporting be on consumer report

Criterion 5: Related and Competing Measures

Related or competing measures

None

Harmonization

N/A

Pre-meeting public and member comments

Identifying person- and family-centered (PFCC) quality measures for home and community-based services (HCBS) is important, especially in developing accountability for the person-centered care requirements in the Centers for Medicare & Medicaid Services HCBS regulations. PFCC quality measures for HCBS are also becoming increasingly important as health care and long-term services and supports become integrated. The HCBS Experience of Care measures collect information from the perspective of the individual, and as such have a person-centered focus. After reviewing the survey questions to be included for the HCBS measure, The SCAN Foundation (Foundation) recommends adjusting or removing the following questions.

Staff listen and communicate well

Survey items 29 and 42 identified as part of the outcome measure for staff listening and communicating well is phrased, “How often are the explanations [personal assistance/behavioral health staff] or [homemaker] gives you hard to understand because of an accent or the way he or she speaks English?” While it is important to identify whether communication between the personal assistance/behavioral health staff/homemaker and the individual receiving services is clearly understood, the way this question is phrased does not effectively address cultural competencies and potential language barriers as it assumes the person receiving care is a native English speaker. The Foundation suggests reframing or removing survey items 29 and 42 to capture whether someone is generally able to understand the provider, spoken to in a language they understand, and can effectively communicate instructions, wishes, and concerns with staff. We acknowledge that survey item 31, “How often do [personal assistance/behavioral health staff] explain things in a way that is easy to understand?” may already addresses the communication concern effectively.

Physical safety measure

The Foundation applauds the inclusion of measures addressing physical safety. However, the proposed measure, “Do any staff that you have now hit you or hurt you?” included in isolation raises concerns. The survey question does not clearly identify new accounts of abuse as opposed to reports that have been addressed and does not appear to include follow up questions for to help with addressing any current concerns. If this measure is to be included, we recommend including additional questions to better understand the current situation in the event of an affirmative response and a clear protocol outlining how to the surveyor should respond to ensure the individual’s safety.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Home and Community Based Services (HCBS) Experience of Care (EoC) Survey
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: Click here to enter a date

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☒Patient-reported outcome (PRO): Experience with Care

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

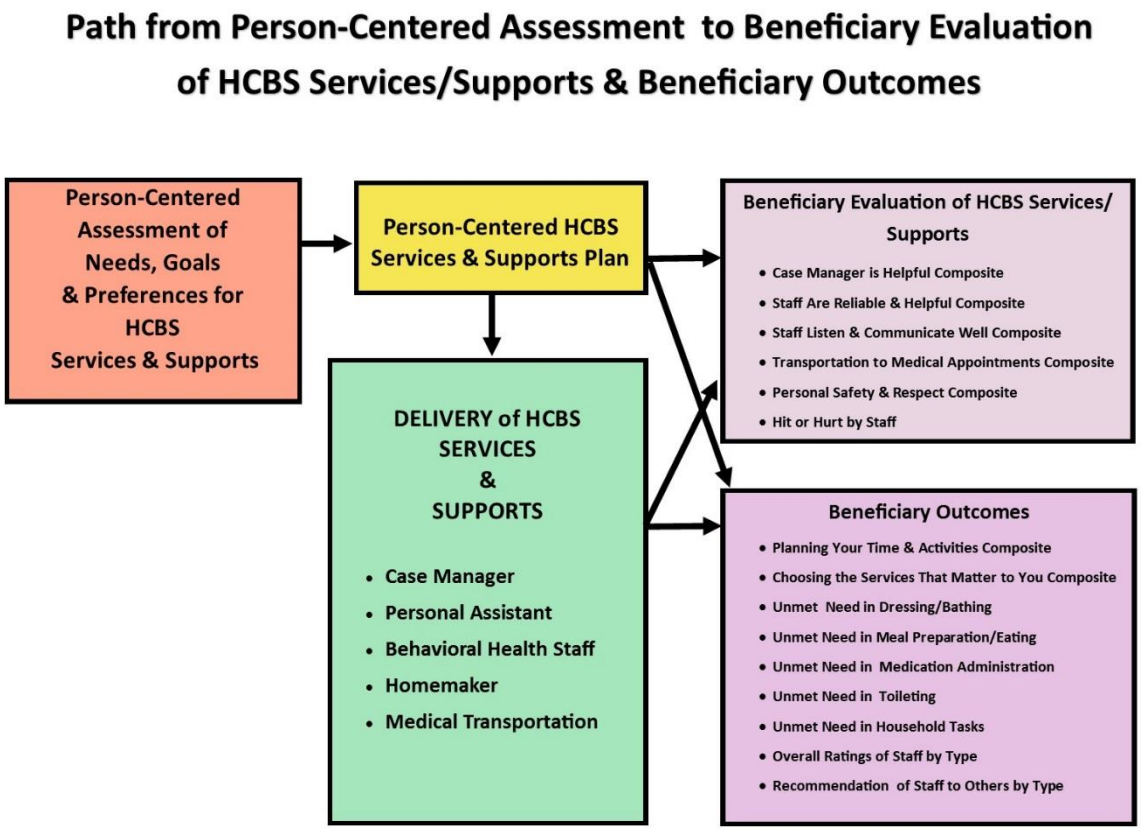
☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

The following diagram illustrates the path to the beneficiary outcomes proposed in this submission, starting with the key processes (i.e., person-centered assessment and service planning) and resulting services (i.e., HCBS services and supports) that are expected to influence the beneficiary assessment of services/supports as well as beneficiary outcomes.



1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

The person-centered approach to beneficiary assessment (of need, goals and preferences), together with person-centered service planning, is expected to influence -- either directly or indirectly via service delivery -- both beneficiary evaluation of services/supports as well as beneficiary outcomes.

The person-centered approach that drives and shapes the beneficiary experience is a fundamental tenet of Medicaid HCBS programs. In 2014, CMS issued new regulations that require Medicaid HCBS programs to work with beneficiaries to develop a person-centered service plan that (a) has individually identified goals and preferences to assist the person in achieving personally-identified outcomes and (b) insures the delivery of services/support in a manner that reflect personal preferences and choice.[1, 2, 3]

The person-centered service planning process is expected to directly influence three composite outcome measures in the following ways:

- A primary case manager responsibility is working with the beneficiary to develop a services/supports plan which in turn will determine the services/supports that the beneficiary receives. Once the services/supports plan has been developed, the case manager also has responsibility for monitoring the plan's implementation to insure it meets the beneficiary's needs/preferences and supports the person in achieving their goals. Thus, it is expected that the case manager's role in both the service planning process and service monitoring will affect the beneficiary's evaluation of the case manager as captured in the composite measure "*Case Manager is Helpful."*
- The purpose of Medicaid HCBS programs is not merely to provide a service(s) but to support beneficiaries' ability to live as they want in the community. Thus, the person-centered planning process is intended to identify the assistance that the beneficiary requires to direct their own lives, as represented in the outcome measure *"Planning Your Time and Activities."*
- The service planning process is expected to directly affect the composite *"Choosing the Services That Matter To You"* because a fundamental principle of that process is to work with the beneficiary to identify the services of their choosing.

The person-centered service planning processs is expected to indirectly affect beneficiary evalustion of services/supports as a result of whether HCBS providers deliver services and supports in accordance with the plan. These impacts are captured by nearly all beneficiary outcomes (except the composite *"Choosing Services That Matter To You"*).

The delivery of HCBS services/supports by providers is expected to directly impact both beneficiary evaluation of service provision as well as beneficiary outcomes. While there are many types of HCBS services and supports, beneficiary experience with those most commonly delivered to people in Medicaid HCBS programs is the focus of the beneficiary evaluation of service/support-related measures. These most common services and supports include:

- Personal Attendant and Behavioral Health Staff who provide assistance with personal care activities.
- Homemakers who assist beneficiaries in activities such as housekeeping, meal preparation and laundry.
- Case Managers who assess the beneficiary's need for services/supports; work with them to develop a service plan responsive to the person's needs, goals and person preferences; monitor service delivery; and assist the person in arranging more/different services as their needs and circumstances change.
- Medical Transportation which provides transportation to medical appointments.

The delivery of these HCBS services/supports is expected to mitigate beneficiary unmet needs as well as influence how beneficiaries assess their experience with the provision of services/supports. The delivery of services/supports in a person-centered manner and responsive to beneficiary preferences is also expected to impact the person's assessment

of the degree to which they have control over planning their daily activities (as measured by the composite *"Planning Your Time and Activities"*).

References

Guidance to HHS Agencies for Implementing Principles of Section 2402(a) of the Affordable Care Act: Standards for Person-Centered Planning and Self-Direction in Home and Community-Based Services Programs: http://www.acl.gov/Programs/CIP/OCASD/docs/2402-a-Guidance.pdf

[2] 2016 Medicaid HCBS Rule in Federal Register: https://federalregister.gov/a/2014-00487

[3] CMS Fact Sheet on 2014 Medicaid HCBS Rule: https://www.medicaid.gov/medicaid-chip-program-information/by-topics/long-term-services-and-supports/home-and-community-based-services/downloads/final-rule-fact-sheet.pdf

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to subcriterion 1b).

Brief Measure Information

NQF #: 2967
De.2. **Measure Title:** Home and Community Based Services (HCBS) Experience of Care (EoC) Measures
Co.1.1. **Measure Steward:** Centers for Medicare and Medicaid Services
De.3. **Brief Description of Measure:** Home and Community Based Services (HCBS) Experience of Care (EoC) measures derive from a cross disability survey to elicit feedback from adult Medicaid beneficiaries receiving home and community based services (HCBS) about the quality of the long-term services and supports they receive in the community.

The measures consist of seven scale measures, 6 global rating and recommendation measures and 6 individual measures:
Scale Measures
1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Ratings Measures
8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendations Measures
11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends –– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)
1b.1. **Developer Rationale:** Scale Measures
Staff are Reliable and Helpful. Assessing the performance of Medicaid direct care providers (i.e., personal assistants, behavioral health staff, homemakers) from the perspective of the beneficiary is important in evaluating the quality of services they render. This measure is based on beneficiary assessment of direct care staff reliability (showing up on time, stay as long as supposed to, communicate absences) and sensitivity to their privacy needs during the provision of personal care.

Staff Listen and Communicate Well. This measure is based on beneficiary assessment of direct care staff's communication skills and responsiveness to the person's needs. Specifically communication in a way that is understood by the beneficiary, respectful, and staff who listen carefully to what the beneficiary needs/wants and who, therefore, understand what the beneficiary needs. This is essential to the delivery of person-centered care and support. Person-centered care and support is required in Medicaid HCBS programs (Federal Register: https://federalregister.gov/a/2014-00487).

Case Manager Is Helpful. In HCBS programs, the case manager is responsible for monitoring the beneficiary's receipt of services and supports to ensure the service plan is being implemented as specified and that the person's needs are being adequately met. In order to meet these requirements, the case manager must be available to the beneficiary when s/he contacts him/her, and responsive to their changing/emerging needs. This measure is based on the beneficiary's assessment of case manager accessibility and responsiveness.

Choosing Services That Matter to You. A basic tenet of Medicaid HCBS services is that the beneficiary is involved in choosing their services/supports so that the service plan is truly person-centered, and that direct care staff implement the service plan in a person-centered manner. This measure is based on the beneficiary's assessment of the extent to which their service plan and direct care workers are person-centered.

Transportation to Medical Appointments. The health and welfare of beneficiaries must be ensured in the delivery of Medicaid HCBS (42 CFR §441: 302). Integral to assuring the health of beneficiaries is getting to medical appointments. This composite is based on the beneficiary's assessment of the extent to which they have transportation to medical appointments, whether the transportation provider is reliable, and whether the transportation is sufficiently accessible.

Planning your time and activities. Medicaid home and community-based services and supports should facilitate outcomes that are consistent with allowing beneficiaries to live the lives they choose – both in terms of daily routine as well as socializing with family and friends, and engaging in community activities. This measure is based on the beneficiary's assessment of the extent to which they have choice and control over these aspects of their lives.

Personal Safety and Respect. Beneficiaries of Medicaid HCBS should be assured that HCBS providers treat them with respect, that they will not be financially exploited by providers coming into their homes, and that they have someone to go to if they are treated badly. This measure will help HCBS programs assess this aspect of program quality.

Individual Item Measures

Global Ratings of Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) – separate measures per staff type. In concert with more specific measures and scale measures, global ratings provide additional information for assessing program quality and can be used as a metric in evaluating quality improvement.

Would Recommend Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) to Family and Friends –separate measures per staff type. Beneficiaries' recommendation are yet another aspect of global experience with a program, and can be used for evaluating program quality and in quality improvement initiatives.

Individual Unmet Need Measures:

- Unmet Need in Dressing/bathing Due to Lack of Help
- Unmet Need in Meal Preparation/Eating Due to Lack of Help
- Unmet Need in Medication Administration Due to Lack of Help
- Unmet Need in Toileting Due to Lack of Help
- Unmet Need with Household Tasks Due to Lack of Help

None of the Unmet Need items were captured in a scale measure because they did not correlate with each other in factor analysis. But the advisory panel for the measures development strongly recommended all unmet need stand-alone items be treated as individual measures as the evaluation of unmet need in HCBS is critically important for determining program quality. One of the most basic reasons for the existence of HCBS programs is to meet self-care needs (bathing, dressing, toileting, medication administration) and needs that, if not met, make successful community living untenable (meal preparation/eating, cleaning/laundry). These measures are intended for use in assessing program quality and for quality improvement initiatives.

Hit or Hurt by Staff. This item was not retained in the Personal Safety and Respect scale measure due to low variation within responses. However, the advisory panel for the measures development felt this measure is important for establishing the personal safety of program beneficiaries, as physical abuse by staff is a "never event" that should be tracked in any HCBS quality management system.

S.4. **Numerator Statement:** HCBS service experience is measured in the following areas. Attached Excel Table S.2b includes the specific item wording for each measure and the response options that go into the numerator.

Scale Measures

1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Rating Measures

8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendation Measures

11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends — average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)
S.7. **Denominator Statement:** The denominator for all measures is the number of survey respondents. Individuals eligible for the HCBS survey include Medicaid beneficiaries who are at least 18 years of age in the sample period, and have received HCBS services for 3 months or longer. Eligibility is further determined using three cognitive screening items, administered during the interview:

Q1. Does someone come into your home to help you? (Yes, No)
Q2. How do they help you?
Q3. What do you call them?

Individuals who are unable to answer these cognitive screening items are excluded. Some measures also have topic-specific screening items as well. Additional detail is provided in S.9.
S.10. **Denominator Exclusions:** Individuals less than 18 years of age and individuals that have not received HCBS services for at least 3 months should be excluded. During survey administration, additional exclusions include individuals that failed any of the cognitive screening items mentioned in the denominator statement below.

De.1. **Measure Type:** PRO
S.23. **Data Source:** Patient Reported Data/Survey
S.26. **Level of Analysis:** Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
HCBS_EoC_NQF_Measures_evidence-attachment_3-29-2016.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

Scale Measures

Staff are Reliable and Helpful. Assessing the performance of Medicaid direct care providers (i.e., personal assistants, behavioral health staff, homemakers) from the perspective of the beneficiary is important in evaluating the quality of services they render. This measure is based on beneficiary assessment of direct care staff reliability (showing up on time, stay as long as supposed to, communicate absences) and sensitivity to their privacy needs during the provision of personal care.

Staff Listen and Communicate Well. This measure is based on beneficiary assessment of direct care staff's communication skills and responsiveness to the person's needs. Specifically communication in a way that is understood by the beneficiary, respectful, and staff who listen carefully to what the beneficiary needs/wants and who, therefore, understand what the beneficiary needs. This is essential to the delivery of person-centered care and support. Person-centered care and support is required in Medicaid HCBS programs (Federal Register: https://federalregister.gov/a/2014-00487).

Case Manager Is Helpful. In HCBS programs, the case manager is responsible for monitoring the beneficiary's receipt of services and supports to ensure the service plan is being implemented as specified and that the person's needs are being adequately met. In order to meet these requirements, the case manager must be available to the beneficiary when s/he contacts him/her, and responsive to their changing/emerging needs. This measure is based on the beneficiary's assessment of case manager accessibility and responsiveness.

Choosing Services That Matter to You. A basic tenet of Medicaid HCBS services is that the beneficiary is involved in choosing their services/supports so that the service plan is truly person-centered, and that direct care staff implement the service plan in a person-centered manner. This measure is based on the beneficiary's assessment of the extent to which their service plan and direct care workers are person-centered.

Transportation to Medical Appointments. The health and welfare of beneficiaries must be ensured in the delivery of Medicaid HCBS (42 CFR §441: 302). Integral to assuring the health of beneficiaries is getting to medical appointments. This composite is based on the beneficiary's assessment of the extent to which they have transportation to medical appointments, whether the transportation provider is reliable, and whether the transportation is sufficiently accessible.

Planning your time and activities. Medicaid home and community-based services and supports should facilitate outcomes that are consistent with allowing beneficiaries to live the lives they choose – both in terms of daily routine as well as socializing with family and friends, and engaging in community activities. This measure is based on the beneficiary's assessment of the extent to which they have choice and control over these aspects of their lives.

Personal Safety and Respect. Beneficiaries of Medicaid HCBS should be assured that HCBS providers treat them with respect, that they will not be financially exploited by providers coming into their homes, and that they have someone to go to if they are treated badly. This measure will help HCBS programs assess this aspect of program quality.

Individual Item Measures
Global Ratings of Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) – separate measures per staff type. In concert with more specific measures and scale measures, global ratings provide additional information for assessing program quality and can be used as a metric in evaluating quality improvement.

Would Recommend Staff (i.e., Personal Assistance/Behavioral Health Staff, Homemaker, Case Manager) to Family and Friends –separate measures per staff type. Beneficiaries' recommendation are yet another aspect of global experience with a program, and can be used for evaluating program quality and in quality improvement initiatives.

Individual Unmet Need Measures:
- Unmet Need in Dressing/bathing Due to Lack of Help
- Unmet Need in Meal Preparation/Eating Due to Lack of Help
- Unmet Need in Medication Administration Due to Lack of Help
- Unmet Need in Toileting Due to Lack of Help
- Unmet Need with Household Tasks Due to Lack of Help

None of the Unmet Need items were captured in a scale measure because they did not correlate with each other in factor analysis. But the advisory panel for the measures development strongly recommended all unmet need stand-alone items be treated as individual measures as the evaluation of unmet need in HCBS is critically important for determining program quality. One of the most basic reasons for the existence of HCBS programs is to meet self-care needs (bathing, dressing, toileting, medication administration) and needs that, if not met, make successful community living untenable (meal preparation/eating, cleaning/laundry). These measures are intended for use in assessing program quality and for quality improvement initiatives.

Hit or Hurt by Staff. This item was not retained in the Personal Safety and Respect scale measure due to low variation within responses. However, the advisory panel for the measures development felt this measure is important for establishing the personal safety of program beneficiaries, as physical abuse by staff is a "never event" that should be tracked in any HCBS quality management system.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See tables 1b.2a and 1b.2b in attached tables.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not applicable.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See tables 1b.4a, 1b.4b, 1b.4c, and 1b.4d in attached tables.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

The measures in this submission focus on people who are elderly with disabilities, individuals with physical disabilities, persons with intellectual/developmental disability, individuals with brain injury, and those with serious mental illness. who receive Medicaid-funded home and community-based services. The Medicaid population with disabilities is, by definition, a population with substantially limited economic resources. Consistent with Medicaid status, adults with disability have a higher poverty rate than those without disability [age 18-64: 28.2% vs. 13.9%, respectively; age 65+: 13.0% vs 7.5% respectively (U.S. Census Bureau, 2014a)]. In addition, U.S. working age adults (Age 18-64) with disability have a lower employment rate than their non-disabled peers [34.4% vs. 75.4% (U.S. Census Bureau, 2014b)].

In terms of racial/ethnic disparities, Blacks, Hispanics and American Indians/Alaskan Natives (AIAN) have higher prevalence of disabilities in self-care and independent living than does the total U.S. adult population with these types of disabilities. These types of disabilities mirror those that beneficiaries in Medicaid HCBS programs tend to exhibit. In the US, 2.1% of the adult population has self-care disabilities and 6.1% have independent living disabilities, respectively. This contrasts to Blacks with respective prevalence of 5.7% and 9.2%; Hispanics at 4.8% and 7.7%; and AIAN at 6.6% and 11.4% (CDC, 2013).

Safety is a major concern for programs serving people with disabilities, who experience higher rates of violent crime victimization. The rate of victimization from violent crime for the U.S. population without disabilities is 14 per 1,000 population. For people with disabilities of the type served in HCBS programs (i.e., disabilities in self-care and independent living), the rates are 26.0/1,000 and 32.4/1,000, respectively. Of most relevance to the safety-related measures in this submission would be statistics on victimization from abuse by paid caregivers; however, the Department of Justice's estimates do not identify paid caregivers as a category of perpetrator (Harrell, 2015).

U.S. Census Bureau. (2014). 2014 American Community Survey, 1-Year Estimates, American FactFinder, Table B18130; http://factfinder.census.gov .

U.S. Census Bureau. (2014). 2014 American Community Survey, 1-Year Estimates, American FactFinder, Table B18120; http://factfinder.census.gov .

Center for Disease Control, Online Disability and Health Data System. (2013). http://www.cdc.gov/ncbddd/disabilityandhealth/dhds.html. Data from the 2013 Behavioral Risk Factor Surveillance System (BRFSS).

Harrell, E. (2015). Crime Against Persons with Disabilities, Statistical Tables. U.S. Department of Justice, May 2015, http://www.bjs.gov/content/pub/pdf/capd0913st.pdf.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

High resource use

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

The development and testing of the measures included in this submission are in direct response to the dearth of performance and quality measures for the increasing HCBS population. As pointed out in a recent report from the NQF project on Home and Community-Based Services Quality: "… there is a lack of systematic measurement of the quality of HCBS across payers and delivery systems (NQF, 2015)."

Rigorously tested quality measures for HCBS is becoming increasingly important as government funding for long-term care has shifted from the provision of care in institutional settings to care at home and in the community. For the first time, in 2013, Medicaid expenditures for HCBS surpassed institutional expenditures, and the trend is expected to continue in the years ahead. The amount of state and federal Medicaid expenditures that are devoted to HCBS has steadily increased since the introduction of Medicaid HCBS programs over 35 years ago. In 2013, Medicaid expenditures for HCBS totaled $74.8 billion (Eiken et al., 2013).

Of all Medicaid funding for individuals receiving long-term services and supports (community-based and institutional care), HCBS accounted for 72% of spending in programs targeting people with developmental disabilities, 40% of spending for programs targeting older people and people with physical disabilities, and 36% of spending for programs serving individuals with serious mental illness or serious emotional disorders (Eiken et al., 2013).

An estimated 3.4 million people used Medicaid HCBS in 2011, 71 percent of all LTSS beneficiaries. This figure includes 1,567,198 people who received services authorized under Section 1915(c) of the Social Security Act, commonly referred to as "HCBS waivers" (Eiken et al., 2015). In a separate report focused on HCBS waivers, CMS-approved State Medicaid reports (from the CMS Reporting Form 372) indicated the following number of people served by population in 2012:

- 792,261 were elders or people with physical disabilities;
- 602,958 were persons with intellectual or developmental disabilities;
- 11,547 were persons with serious mental illness or serious emotional disorder; and
- 10,959 were individuals with brain injury (Eiken, 2012).

It should be noted that these statistics are an undercount of the actual number of individuals receiving 1915(c) waiver services in 2012. Data reported by states on the CMS Form 372 Reports represent 284 of the 305 1915(c) waiver programs in operation that year. Only 372 Reports submitted and approved by CMS are represented in the statistics cited above. In addition to the 1915(c) HCBS waiver programs, in 2012 four states provided services/supports to Medicaid beneficiaries through Medicaid managed long-term services and supports (MLTSS) programs authorized under Section 1115 of the Social Security Act. The numbers served in these MLTSS programs is not available from the 372 Reports.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

National Quality Forum. (2015). Addressing Performance Measure Gaps in Home and Community-Based Services to Support Community Living: Synthesis of Evidence and Environmental Scan, Interim Report. December 18, 2015.

Eiken, S., Sredl, K., Burwell, B., and Saucier, P. (2013). Medicaid Expenditures for Long-Term Services and Supports (LTSS) in FY 2013: Home and Community-Based Services were a Majority of LTSS Spending. Truven Health Analytics, June 30, 2015. https://www.medicaid.gov/medicaid-chip-program-information/by-topics/long-term-services-and-supports/downloads/ltss-expenditures-fy2013.pdf

Eiken, S., Sredl, K., Saucier, P., Burwell, B. (2015). Medicaid Long-Term Services and Supports Beneficiaries in 2011, Truven Health Analytics, September 22, 2015. https://www.medicaid.gov/medicaid-chip-program-information/by-topics/long-term-services-and-supports/downloads/ltss-beneficiaries-report-2011.pdf

Eiken, S. (2012). Medicaid 1915(c) Waiver Data Based on CMS 372 Report, 2011-2012, Truven Health Analytics, September 17, 2015. https://www.medicaid.gov/medicaid-chip-program-information/by-topics/long-term-services-and-supports/downloads/cms-372-report-2012.pdf

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

During the development of the survey from which these measures are, the HCBS beneficiary audience as well as stakeholders in the broader HCBS community have consistently supported the proposed measures as necessary and important.

Truven Health Analytics conducted a literature review for AHRQ that included identifying measures and gaps in the measures. The HCBS EoC Survey team conducted a follow-up literature review for the time period of 2007 through 2010.

The research team received input from a focus group and interviews, and CMS posted 60-day and 30-day Federal Register notices on May 18, 2012 and July 24, 2012, respectively, for public comment on the proposed data collection (as required by the OMB Paperwork Requirement Act). No comments were received.

In addition, there was a Federal Advisory Panel consisting of:

- CMS-Disabled and Elderly Health Programs Group: Anita Yuskauskas (Chair), Mary Sowers, Kathy Poisal, Mary Beth Ribar, Sara Fogler, Carey Appold,
- CMS-Children & Families Health Program Group: Charlie Mackay and John Young
- CMS-Center for Drug and Health Plan Choice: Liz (Elizabeth) Goldstein, Suzanne Rotwein, Lori Teichman, Ted (Edward) Sekscenski, Bill (William) Lehrman, Barb (Barbara) Crawley
- Agency for Healthcare Research and Quality: DEB Potter, Judy Sangl

The research team identified and invited experts and key stakeholders, including representatives of state HCBS programs, self-advocacy groups for people with disabilities, survey development and reporting experts, CAHPS Consortium representatives, and Federal Government staff, to provide feedback on the development of the survey and the field test process. The organizations represented include:

- Linda Anthony, Disability Rights Network of Pennsylvania and ADAPT, Consumer advocate—adults with physical disabilities
- Julie Brown, RAND Corporation, CAHPS Consortium
- Marcus Canaday, West Virginia Bureau for Medical Services, State HCBS programs for adults with physical disabilities
- Steve Dunaway, Florida Agency for Persons with Disabilities, State HCBS programs for adults with intellectual disabilities
- Chester Finn, Self Advocates Becoming Empowered, Consumer advocate—adults with intellectual disabilities
- Michelle Goody, Massachusetts Medicaid, Medicaid
- Ron Honberg and Sita Diehl, National Alliance on Mental Illness, Consumer advocate—adults with mental illness
- Ari Houser, AARP, Consumer advocate—older adults with disabling/chronic conditions
- Christian Koltonski, Colorado Medicaid, Medicaid
- Jeanne Levelle, Louisiana Medicaid, Medicaid
- Ted Lutterman, National Association of State Mental Health Program Directors, State HCBS programs for adults with mental illness
- Chas Moseley and Nancy Thaler, National Association of State Directors of Developmental Disabilities Services, State HCBS programs for adults with intellectual disabilities
- Sue Palsbo, George Mason University, Survey development for people with physical disabilities
- Teresa Richard, Texas Department of Aging and Disability Services, State HCBS programs—all populations

• Steve Staugaitis, University of Massachusetts Medical School, Performance measures for people with intellectual disabilities
• John Thompson and Kelsey Walter, National Association of States United for Aging and Disabilities, State HCBS programs for older adults with disabilities
• Sally Varney, New Hampshire Medicaid, Medicaid
• Sandeep Wadhwa and Matt Salo, National Association of Medicaid Directors (NAMD) and Colorado Department of Health Care Policy and Financing, State HCBS programs—all populations
• Lorraine Wargo, National Association of State Head Injury Administrators, State HCBS programs for adults with head injuries

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Mental Health : Serious Mental Illness, Neurology : Brain Injury, Neurology : Cognitive Impairment/Dementia

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Care Coordination, Functional Status, Health and Functional Status : Development/Wellness, Patient and Family Engagement, Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
The survey and related materials will be available on CMS' Medicaid.gov website; they will also appear on AHRQ's website if the survey receives the CAHPS trademark. The survey instruments in English and Spanish are attached for reference.

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** HCBS_EoC_Supplementary_Tables_3_29_16-635948620440450044.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Not applicable.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

HCBS service experience is measured in the following areas. Attached Excel Table S.2b includes the specific item wording for each measure and the response options that go into the numerator.

Scale Measures
1. Staff are reliable and helpful – average of applicable beneficiary scores on 6 survey items
2. Staff listen and communicate well – average of applicable beneficiary scores on 11 survey items
3. Case manager is helpful - average of applicable beneficiary scores on 3 survey items
4. Choosing the services that matter to you - average of applicable beneficiary scores on 2 survey items
5. Transportation to medical appointments - average of applicable beneficiary scores on 3 survey items
6. Personal safety and respect - average of applicable beneficiary scores on 3 survey items
7. Planning your time and activities - average of applicable beneficiary scores on 6 survey items

Global Rating Measures
8. Global rating of personal assistance and behavioral health staff- average score on a 0-10 scale
9. Global rating of homemaker- average score on a 0-10 scale
10. Global rating of case manager- average score on a 0-10 scale

Recommendation Measures
11. Would recommend personal assistance/behavioral health staff to family and friends – average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
12. Would recommend homemaker to family and friends –– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)
13. Would recommend case manager to family and friends– average score on a 1-4 scale (Definitely no, Probably no, Probably yes, Definitely yes)

Unmet Needs Measures
14. Unmet need in dressing/bathing due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
15. Unmet need in meal preparation/eating due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
16. Unmet need in medication administration due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
17. Unmet need in toileting due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)
18. Unmet need with household tasks due to lack of help–average score on a 1-4 scale (Never, Sometimes, Usually, Always)

Physical Safety Measure
19. Hit or hurt by staff –average score on a 1-4 scale (Never, Sometimes, Usually, Always)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
There is no time reference for measures because cognitive testing showed that this was cognitively burdensome for respondents. This follows the same approach as the CAHPS Nursing Home Long Stay survey, which measures experience of care for a similar population and used a non-specific reference period based on cognitive testing findings (Sangl et al., 2007).

The frequency of data collection/aggregation will be at the discretion of state users, as CMS has determined the survey from which the measures are derived will be conducted on a voluntary by states. It is anticipated that states would

field the survey no more frequently than annually per HCBS program. Some states may choose to field it less frequently than annually. Reporting of measures would follow at intervals paralleling data collection time frames.

The research team wanted to assess whether individuals could respond to a question with a time reference. In the first draft of the survey that was cognitively tested, the team used two approaches, one with a time reference and one without. Round 1 cognitive testing results were inconclusive. We then conducted a second round of cognitive testing with an experiment to evaluate including a time referent versus excluding it. In round 2 of cognitive testing, the team asked nine participants to answer a question with a time reference and one question without a time reference. Once they answered the two questions and related follow up questions, the team asked the participant how the items were different. Consistent with the findings from the cognitive testing of the CAHPS Nursing Home Long Stay Resident instrument, the standard CAHPS 6-month time referent did not test well. A few respondents either could not specify what this time period meant to them, or indicated that there would be no difference in their response based on whether a defined look-back period or the indefinite present were used. Therefore, the team decided to word items in the indefinite present, following the model of the Nursing Home Long Stay instrument targeting a population similar to the HCBS population.

Sangl, J., Buchanan, J., Cosenza, C., Bernard, S., Keller, S., Mitchell, N., and Larwood D. (2007). The development of a CAHPS instrument for Nursing Home Residents (NHCAHPS). J Aging Soc Policy. 19(2):63-82. PubMed PMID: 17409047.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Attached Excel Table S.2b includes the specific item wording for each measure and the response options that go into the numerator.

Scale Measures:
The numerator for each Scale measure includes the number of respondents who chose a substantive answer for at least one item in that scale. Depending on the response option set for the item, substantive answers are considered to be “Never”, “Sometimes”, “Usually”, “Always”, “Mostly Yes”, “Mostly No”, “Yes”, “No”, “None of the things that are important to you”, “Some of the things that are important to you”, “Most of the things that are important to you”, or “All of the things that are important to you”. Item numbers and item text are listed below.

Staff are reliable and helpful – survey items 13 14 15 19 37 38
13: How often do [personal assistance/behavioral health staff] come to work on time?
14: How often do [personal assistance/behavioral health staff] work as long as they are supposed to?
15: Sometimes staff cannot come to work on a day that they are scheduled. When staff cannot come to work on a day that they are scheduled, does someone let you know if [personal assistance/behavioral health staff] cannot come that day?
19: How often do [personal assistance/behavioral health staff] make sure you have enough personal privacy when you dress, take a shower, or bathe?
37: How often do [homemakers] come to work on time?
38: How often do [homemakers] work as long as they are supposed to?

Staff listen and communicate well – survey items 28 29 30 31 32 33 41 42 43 44 45
28: How often are [personal assistance/behavioral health staff] nice and polite to you?
29: How often are the explanations [personal assistance/behavioral health staff] gives you hard to understand because of an accent or the way he or she speaks English?*

30: How often do [personal assistance/behavioral health staff] treat you the way you want them to?
31: How often do [personal assistance/behavioral health staff] explain things in a way that is easy to understand?
32: How often do [personal assistance/behavioral health staff] listen carefully to you?
33: Do you feel [personal assistance/behavioral health staff] know what kind of help you need with everyday activities, like getting ready in the morning, getting groceries, or going places in your community?
41: How often are [homemakers] nice and polite to you?
42: How often are the explanations [homemaker] gives you hard to understand because of an accent or the way the provider speaks English?*
43: How often do [homemakers] treat you the way you want them to?
44: How often do [homemakers] listen carefully to you?
45: Do you feel [homemakers] know what kind of help you need?

Case manager is helpful – survey items 49 51 53
49: Can you contact this [case manager] when you need to?
51: Did this [case manager] work with you when you asked for help with getting or fixing equipment?
53: Did this [case manager] work with you when you asked for help with getting other changes to your services?

Choosing the services that matter to you – survey items 56 57
56: Does your [program-specific term for "service plan"] include . . . ?
57: Do you feel [personal assistance/behavioral health staff] know what's on your [program-specific term for "service plan"], including the things that are important to you?

Transportation to medical appointments – survey items 59 61 62
59: Medical appointments include seeing a doctor, a dentist, a therapist, or someone else who takes care of your health. How often do you have a way to get to your medical appointments?
61: Are you able to get in and out of this ride easily?
62: How often does this ride arrive on time to pick you up?

Personal safety and respect – survey items 64 65 68
64: Is there a person you can talk to if someone hurts you or does something to you that you don't like?
65: Do any of the [personal assistance/behavioral health staff, homemakers, or your case managers] that you have now take your money or your things without asking you first?*
68: Do any [staff] that you have now yell, swear, or curse at you?*

Planning your time and activities – survey items 75 77 78 79 80 81
75: When you want to, how often can you get together with these family members who live nearby?
77: When you want to, how often can you get together with these friends who live nearby?
78: When you want to, how often can you do things in the community that you like?
79: Do you need more help than you get now from [personal assistance/behavioral health staff] to do things in your community?*
80: Do you take part in deciding what you do with your time each day?
81: Do you take part in deciding when you do things each day—for example, deciding when you get up, eat, or go to bed?

Global Ratings Measures:
The numerator for each Global measure includes the number of respondents who chose a substantive answer for that item. Depending on the response option set for the item, substantive answers are considered to be a 0-10 rating, "Excellent", "Very good", "Good", "Fair" or "Poor". Item numbers and item text are listed below.

Global rating of personal assistance and behavioral health staff– survey item 35

35: Using any number from 0 to 10, where 0 is the worst help from {personal assistance/behavioral health staff} possible and 10 is the best help from {personal assistance/behavioral health staff} possible, what number would you use to rate the help you get from {personal assistance/behavioral health staff}?

Global rating of homemaker – survey item 46
46: Using any number from 0 to 10, where 0 is the worst help from {homemakers} possible and 10 is the best help from {homemakers} possible, what number would you use to rate the help you get from {homemakers}?

Global rating of case manager– survey item 54
54: Using any number from 0 to 10, where 0 is the worst help from {case manager} possible and 10 is the best help from {case manager}possible, what number would you use to rate the help you get from {case manager}?

Recommendation Measures:
The numerator for each Recommendation measure includes the number of respondents who chose "Definitely no", "Probably no", "Probably yes", "Definitely yes". Item numbers and item text are listed below.

Would recommend personal assistance/behavioral health staff to family and friends – survey item 36
36: Would you recommend the {personal assistance/behavioral health staff} who help you to your family and friends if they needed help with everyday activities? Would you say you recommend the {personal assistance/behavioral health staff} . . .

Would recommend homemaker to family and friends – survey item 47
47: Would you recommend the {homemakers} who help you to your family and friends if they needed {program-specific term for homemaker services}? Would you say you recommend the {homemakers} . . .

Would recommend case manager to family and friends– survey item 55
55: Would you recommend the {case manager} who helps you to your family and friends if they needed {program-specific term for case-management services}? Would you say you recommend the {case manager} . . .

Unmet Needs Measures:
The numerator for each Unmet Needs measure includes the number of respondents who answered "yes" for that item. Item numbers and item text are listed below.

Unmet need in dressing/bathing due to lack of help - survey item 18
18: Is this because there are no {personal assistance/behavioral health staff} to help you?

Unmet need in meal preparation/eating due to lack of help - survey item 22
22: Is this because there are no {personal assistance/behavioral health staff} to help you?

Unmet need in medication administration due to lack of help - survey item 25
25: Is this because there are no {personal assistance/behavioral health staff} to help you?

Unmet need in toileting due to lack of help - survey item 27
27: Do you get all the help you need with toileting from {personal assistance/behavioral health staff} when you need it?

Unmet need with household tasks due to lack of help - survey item 40
40: Is this because there are no {homemakers} to help you? [ASK IF HOMEMAKER IS THE SAME AS PCA STAFF]

Physical Safety Measure:
The numerator for the following Physical Safety measure includes the number of respondents who answered "yes" for this item. The item number and item text is listed below.

Hit or hurt by staff – survey item 71
71: Do any {staff} that you have now hit you or hurt you?

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator for all measures is the number of survey respondents. Individuals eligible for the HCBS survey include Medicaid beneficiaries who are at least 18 years of age in the sample period, and have received HCBS services for 3 months or longer. Eligibility is further determined using three cognitive screening items, administered during the interview:

Q1. Does someone come into your home to help you? (Yes, No)
Q2. How do they help you?
Q3. What do you call them?

Individuals who are unable to answer these cognitive screening items are excluded. Some measures also have topic-specific screening items as well. Additional detail is provided in S.9.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
While there are a myriad of home and community-based services and supports (HCBS) that Medicaid programs provide (at their discretion) to beneficiaries with long-term care needs, the proposed provider-related measures in this submission focus on the most common provider types for adults receiving Medicaid HCBS. These include personal assistance providers, behavioral health staff, homemakers and case managers.

While Medicare-certified home health agencies may provide similar services to Medicare beneficiaries, the Medicare benefit is a post-acute care benefit and typically limited to episodes following hospitalization. Medicaid home and community-based services are a long-term care benefit and support persons with long-term care needs over lengthier durations. Personal assistance services, help in the home by behavioral health staff, and homemaker services typically involve assistance with activities of daily living (bathing, dressing, grooming, toileting, eating; mobility) and instrumental activities of daily living (meal preparation, housework, laundry, food shopping). Case management is an integral component of Medicaid HCBS programs; the role of the case manager includes working with the beneficiary to assesses his/her need for services/supports and to develop a person-centered care/service plan, monitoring service delivery, and responding to the individual's changing needs and circumstances.

Not all HCBS beneficiaries receive all services. Q4, Q6, Q8, and Q11 assess which services the beneficiary receives. Beneficiaries are then eligible for different survey questions based on these responses.

These questions are:
Q4. Do you get {program specific term for personal assistance} at home?
Q6. Do you get {program specific term for behavioral health specialist services} at home?
Q8. Do you get {program specific term for homemaker services} at home?

Q11. Do you get help from {program specific term for case manager services} to help make sure that you have all the services you need?

Scale Measure 1: Staff are reliable and helpful
Q13: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q14: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q15: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q19: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q37: the number of surveys completed by all those who responded "yes" to screener Q8
Q38: the number of surveys completed by all those who responded "yes" to screener Q8

Scale Measure 2: Staff listen and communicate well
Q28: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q29: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q30: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q31: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q32: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q33: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q41: the number of surveys completed by all those who responded "yes" to screener Q8
Q42: the number of surveys completed by all those who responded "yes" to screener Q8
Q43: the number of surveys completed by all those who responded "yes" to screener Q8
Q44: the number of surveys completed by all those who responded "yes" to screener Q8
Q45: the number of surveys completed by all those who responded "yes" to screener Q8

Scale Measure 3: Case manager is helpful
Q49: the number of surveys completed by all those who responded "yes" to screener Q11
Q51: the number of surveys completed by all those who responded "yes" to screener Q11
Q53: the number of surveys completed by all those who responded "yes" to screener Q11

Scale Measure 4: Choosing the services that matter to you
Q56: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q57: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11

Scale Measure 5: Transportation to medical appointments
Q59: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q61: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q62: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11

Scale Measure 6: Personal safety and respect
Q64: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q65: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q68: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11

Scale Measure 7: Planning your time and activities
Q75: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q77: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q78: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q79: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Q80: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11
Q81: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11

Global Rating Measures:
Global rating of personal assistance and behavioral health staff
Q35: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Global rating of homemaker
Q46: the number of surveys completed by all those who responded "yes" to screener Q8
Global rating of case manager
Q54: the number of surveys completed by all those who responded "yes" to screener Q11

Recommendation Measures:
Recommendation of personal assistance and behavioral health staff to family/friends
Q36: the number of surveys completed by all those who responded "yes" to screener Q4 or Q6
Recommendation of homemaker to family/friends
Q47: the number of surveys completed by all those who responded "yes" to screener Q8
Recommendation of case manager to family/friends
Q55: the number of surveys completed by all those who responded "yes" to screener Q11

Unmet Needs Measures:
Unmet need in dressing/bathing due to lack of help -
Q18: the number of surveys completed by all those who responded "yes" to Q17
Unmet need in meal preparation/eating due to lack of help
Q22: the number of surveys completed by all those who responded "yes" to Q21
Unmet need in medication administration due to lack of help
Q25: the number of surveys completed by all those who responded "yes" to Q24
Unmet need in toileting due to lack of help -
Q27: the number of surveys completed by all those who responded "yes" to Q26
Unmet need with household tasks due to lack of help
Q40: the number of surveys completed by all those who responded "yes" to Q39

Personal Safety Measures:
Hit or hurt by staff
Q71: the number of surveys completed by all those who responded "yes" to screener Q4, Q6, Q8, or Q11

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Individuals less than 18 years of age and individuals that have not received HCBS services for at least 3 months should be excluded. During survey administration, additional exclusions include individuals that failed any of the cognitive screening items mentioned in the denominator statement below.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Individuals who are unable to answer one or more of the following cognitive screening items should be excluded. If the respondent is not able to answer (e.g., provides an invalid/nonsensical response, does not respond, or indicates "I don't know"), the interviewer should end the interview.

1. Does someone come into your home to help you? (Yes, No)

2. How do they help you? (open ended)
Examples of correct responses include:
• "Helps me get ready every day"

• "Cleans my home"
• "Works with me at my job"
• "Helps me to do things"
• "Drives me around"

3. What do you call them? (open ended)
Examples of sufficient responses include:
• "My worker"
• "My assistant"
• Names of staff ("Jo", "Dawn", etc.)

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
The intended primary unit of analysis is the Medicaid HCBS program. However, states may wish to stratify by sub-state agencies such as counties or regional entities with program operational and budgetary authority. In some instances, a state may wish to stratify by case-management agency as well, given they are typically viewed as having substantial responsibility for developing beneficiary service and support plans as well as monitoring whether the service/support plan addresses the person's needs and meet their goals.

States are increasingly moving users of Medicaid long-term services and supports, including HCBS, into managed care arrangements (typically referred to as Managed Long-Term Services and Supports or MLTSS) where the managed care organization (MCO) is the primary accountable entity for ensuring HCBS beneficiary, health, welfare and quality of life. As such, we also anticipate some states may want to stratify based on (MCO).

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Case-mix adjustment is done via regression methodology or a covariance adjustment. We use case-mix adjustment to adjust scores for various patient and survey mode characteristics. The research team suggests general health rating, mental health rating, age, gender, whether respondent lives alone, and response option as case- mix adjusters for the HCBS EoC measures based on our analysis. We also recommend including survey mode as an additional adjustment variable and proxy status if proxy respondents are utilized.

The specific survey items used to develop case mix adjustment are:

82. In general, how would you rate your overall health? Would you say . . .
Excellent,
Very good,
Good,
Fair, or
Poor?
DON'T KNOW
REFUSED
UNCLEAR RESPONSE

83. In general, how would you rate your overall mental or emotional health? Would you say . . .
Excellent,
Very good,
Good,
Fair, or
Poor?
DON'T KNOW
REFUSED
UNCLEAR RESPONSE

84. What is your age?
18 TO 24 YEARS GO TO Q85
25 TO 34 YEARS GO TO Q85
35 TO 44 YEARS GO TO Q85
45 TO 54 YEARS GO TO Q85
55 TO 64 YEARS GO TO Q85
65 TO 74 YEARS GO TO Q85
75 YEARS OR OLDER GO TO Q85
DON'T KNOW
REFUSED? GO TO Q85
UNCLEAR RESPONSE

85. Are you male or female?
MALE
FEMALE
DON'T KNOW
REFUSED
UNCLEAR RESPONSE

93. How many adults live at your home, including you?
1 [JUST THE RESPONDENT] ? END SURVEY
2 TO 3
4 OR MORE
DON'T KNOW
REFUSED
UNCLEAR RESPONSE

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
The research team used the CAHPS SAS analysis program to produce the scores which allows users to specify case-mix adjusters. For case-mix adjustment specifications, see pages 54-60 of the Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1 available at https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/2015-Instructions-for-Analyzing-Data-from-CAHPS-Surveys.pdf .

S.16. **Type of score:**
Other (specify):

If other: Case-mix adjusted means

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Scoring specifications for the measures will follow the same general scoring approach as used by other CAHPS surveys that use the CAHPS analysis program. The measures are based on case-mix adjusted means that are transformed into a 0–100 metric. The research team suggests general health rating, mental health rating, age, gender, whether respondent lives alone, and response option as case- mix adjusters for these measures. We also recommend including survey mode as an additional adjustment variable and proxy status if proxy responses are permitted. More information about case-mix adjustment is available in Instructions for Analyzing Data from CAHPS Surveys (available at https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/2015_instructions_for_analyzing_data.pdf).

To create scores for each scale measure:
1. Calculate the case-mix adjusted mean separately for each item in each scale. This process creates the arithmetic mean for each item and adjusts for respondent characteristics identified in the case mix analysis. This makes it more likely that reported differences are due to real differences in program performance, rather than differences in the characteristics of service recipients.
 a. The steps for user-defined calculations of risk-adjusted scores can be found in Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1 available at https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf

2. Calculate the average of the case-mix adjusted means across the items in each scale; use equal weighing of the items.

3. Transform the average from Step 2 to a 0–100 scale as follows: score = [(x – a)/(b – a)]*100, where x = the case-mix adjusted mean from step 1; a = minimum possible value of x; and b = maximum possible value of x. This transformation allows the presentation of different survey-based measures on a common metric.

To create scores for each global rating and individual item measure:
1. Calculate the case-mix adjusted mean for the item. This process is the same as step 1 above. This process creates the arithmetic mean for each item and adjusts for respondent characteristics identified in the case mix analysis.
 a. The steps for user-defined calculations of risk-adjusted scores can be found in Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1 available at https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf

2. Transform the item 0–100 scale (use the same formula as described in Step 2 for calculation of global scale measures).

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

Defining the Sample Frame: Eligibility Guidelines
The intended sample for the HCBS EOC survey that the measures are based on is adult Medicaid beneficiaries age 18 or older who have received HCBS services for 3 months or longer from the intended survey administration. Sampling should be stratified by HCBS program within each state, in order to all comparisons of measure results for each HCBS program to the state mean. The source of the sample frame will be the state Medicaid agency or an entity delegated by the state Medicaid agency (e.g., state agency other than the Medicaid agency that operates the program, a MCO, a case management agency, state county, etc.).

Recommended Number of Completed Surveys
In order to determine the size of the sample, each state should take into account the effective sample size and response rates from the field test. The effective sample size is the number of completed responses needed to obtain a reasonable level of reliability. The research team conducted a pilot test and a field test of the measures with 26 Medicaid HCBS programs across ten states from October 2013 to March 2015. Results suggest that the effective sample size should be 400 people per stratum (with smaller programs including the census). From field test data, we know that the total response rate was 22.0% and this ranged from 9.8% – 31.1% for HCBS programs and modes of administration. Some states may expect a higher response rate in future administrations because of better outreach, pre-survey communications with potential respondents, as well as use of proxies and can adjust their estimated response rate based on these additional considerations.

Proxy Responses
Proxy responses were permitted for the field test of the measures; however, with the exception of the response rate calculations, the analyses described in this report exclude proxy responses. Due to the fact that the proxy data were not collected consistently across states and programs, the research team cannot reliably make inferences about differences between proxy respondents and non-proxy respondents for the field test. Proxy here is defined as anyone who provided help to the beneficiary completing the survey. We do expect states to allow proxy responses in future data collection efforts. Most immediately, TEFT grantees who are implementing the survey instrument will have the option of allowing respondents to receive assistance or to have a proxy. They will receive information about considerations and possible approaches to incorporating proxies in data collection. It will be their decision whether and how to incorporate proxies.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Survey Administration Mode
Due to the impairments (i.e., cognitive, hearing) prevalent among individuals served by HCBS programs, stakeholders recommend that the survey be conducted through in-person interviews. However, the CAHPS consortium urged the research team to assess both in-person as well as phone administration modes. Based on field test results, administering the survey by phone was found appropriate if a statistical adjustment for survey mode is made for mixed-mode administrations. For programs using the survey measures to monitor trends, we recommend not switching modes across survey fielding periods. A mail survey is not recommended for the HCBS population due to the prevalence of cognitive disabilities

Survey Response Options
Based on findings from cognitive testing as well as an experiment conducted as part of the field test, a simplified response option of Mostly Yes/ Mostly No was determined more accessible for some respondents than the standard CAHPS response option of Never/ Sometimes/ Usually/ Always. For the field test, within each mode (Computer-assisted telephone interviewing and Computer-assisted personal interviewing), equal numbers of participants were randomly assigned to one of the two response option formats—either the 4-point response option or the 2 point binary response option. Participants assigned to the standard response option were switched to the simplified response option if they had difficulty responding using these cognitively more challenging options. “Difficulty” was

determined by how well respondents answered the first three survey questions under Getting Needed Services from Personal Assistant and Behavioral Health Staff . If they were unable to answer the questions or had difficulty answering them, the interviewer switched to the alternative format, similar to the CAHPS Nursing Home Long-Stay Resident method.

The interviewer will need to make the determination as to when to use the alternate response option using the following process. If the respondent is unable to respond using the responses "Never, Sometimes, Usually, And Always" as indicated non-verbally or verbally by stating "I don't understand", "I am not sure of the difference" or a similar response, the interviewer should reread the question providing the "Mostly Yes And Mostly No" response option. For the following question, the interviewer should provide the standard responses "Never, Sometimes, Usually, And Always" again, providing the alternate responses of "mostly yes and mostly no" only if the respondent is unable to respond using the standard response. After three unsuccessful attempts to use the standard response, the interviewer should switch to the alternate response and use it throughout the remaining interview.

Including both response modes will allow more respondents to respond to the survey, including individuals with a developmental disability, intellectual/cognitive impairment, or a traumatic brain injury. In cases where both responses are included, the data from the simplified response should be transformed (mostly yes = always and mostly no= never) and pooled with the standard responses for reporting. It is critical to case mix adjust for survey response if both options are offered.

Survey Administration

At least one week prior to survey administration, the states should mail a pre-notification letter on state letterhead to all sampled members, alerting them to expect a phone call about the interview and assuring the sampled members that the survey is endorsed by the state. After the pre-notification letters are mailed, the survey vendors should begin telephone contact of HCBS program participants to introduce the survey, explain the survey's purpose, and schedule the interview date and time. To solicit participation, survey vendors should make at least five call attempts to sampled participants during different call days/times—calling in daytime hours during the week, in the evening, and once on the weekend.

Response Rates

The total response rate was 22.0% from the field test and this ranged from 9.8% – 31.1% for the different HCBS programs. Some states may expect a higher response rate in future administrations because of better outreach, upfront communications, and use of proxies.

The research team calculated the response rate using the American Association for Public Opinion Research (AAPOR) response rate #3 (RR#3):

$I/((I+P) + (R+NC+O) + e(UH+UO))$

Where:

I = complete interviews (3,226)
P = partial interviews (33)
R = refusals and breakoffs (2,442)
NC = noncontact (3,014)
O = other (3,200)
UH = unknown household (3,868)
UO = unknown other (123)
e = estimated proportion of cases of unknown eligibility that are eligible (0.68)

AAPOR defines several options for calculating response rate. Based on the research team's sampling approach, the formula that is most appropriate for these data was RR#3 (http://www.aapor.org/AAPORKentico/Communications/AAPOR-Journals/Standard-Definitions.aspx). The response

rate is the total number of completed surveys divided by the total number of eligible sampled individuals. Households with nonworking or wrong numbers are excluded from the denominator. In some cases, eligibility cannot be determined. For these individuals, RR#3 adjusts the response rate assuming that the rate of response for undetermined households would be the same as the response rate where eligibility could be determined. This is shown in the formula where the number of unknowns (UH + UO) is multiplied by the estimated proportion of cases of unknown eligibility that are eligible (e). The result is a slight upward adjustment of the response rate. Thus, the overall response rate was 21.1 percent (22.3 percent in-person and 20.9 percent for phone).

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data are not imputed for unadjusted scores. Measure scores are calculated at the unit level (e.g. HCBS program) using all available data for individual items. Means for individual survey items are computed individually. These are then averaged across items to calculate the scale measure scores. Therefore, a case with usable data for only some individual survey items can be used in the calculation of scale measure scores for a program. However, only "complete" survey responses (those that answered at least half of key items) are included in all measures calculations.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Patient Reported Data/Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Home and Community Based Services (HCBS) Experience of Care (EoC) Survey
In-person and phone
English and Spanish

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other
If other: Home and Community-Based Services Program

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
HCBS_EoC_NQF_Measures_testing-attachment_3_29_16.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Home and Community Based Services (HCBS) Experience of Care (EoC) Measures
Date of Submission: 3/31/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: HCBS EoC Survey Data*	☒ other: HCBS EoC Survey Data

*Metrics presented throughout are derived from analysis of the Home and Community Based Services Experience of Care Survey funded by the Centers for Medicare and Medicaid Services

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Not applicable

1.3. What are the dates of the data used in testing? October 2013 – March 2015

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: Medicaid HCBS programs	☒ other: Medicaid HCBS programs

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The measured entity is Medicaid HCBS programs. HCBS is a set of services a person receives. Survey responses are compiled to develop scale measures that assess the quality of HCBS services at the program level.

The research team conducted a pilot test and a field test of the survey with 26 Medicaid HCBS programs across ten states. The ten states were geographically dispersed and included AZ, CO, CT, GA, KY, LA, MD, MN, NH, and TN; these states (with the exception of TN) were CMS Testing Experience and Functional Tools (TEFT) Demonstration grantees. These 26 HCBS programs serve a wide array of people including people who are elderly with disabilities, individuals with physical disabilities, persons with intellectual/developmental disability, individuals with brain injury, and those with serious mental illness. Combined, these programs served over 138,000 individuals. A random sample of these (n=21,434) HCBS beneficiaries were invited to complete the survey. The complete analytic dataset consists of surveys from 3,223 total respondents. Of these, 2,336 cases were deemed "complete" (over half of all key items were answered) and were used in the reliability analysis

presented here. The number of returned surveys in each program ranges from 0 to 304. One program was not included in analysis because it did not have any returned surveys.

Exhibit 1. States, Populations, Programs, Authorities, and Total Returned Surveys

State	Population Category	HCBS Program	Funding Authority	Number of Total Returned Surveys
Arizona	Elderly/Physically Disabled	Arizona Long Term Care System (ALTCS), Elderly and Physically Disabled expansion	Medicaid 1115 waiver	127
	ID/DD	Arizona Long Term Care System (ALTCS), Developmental Disability	Medicaid 1115 waiver	58
Colorado	Elderly/Physically Disabled	Elderly, Blind, and Disabled Waiver	Medicaid 1915(c) waiver	151
	ID/DD	Supported Living Services Waiver	Medicaid 1915(c) waiver	92
Connecticut	Elderly	Connecticut Home Care Program for Elders	Medicaid 1915(c) waiver	179
	TBI	Acquired Brain Injury Waiver	Medicaid 1915(c) waiver	115
	SMI	Working for Support and Empowerment (WISE) Waiver	Medicaid 1915(c) waiver	81
Georgia	Physically Disabled, TBI	Independent Care Waiver Program	Medicaid 1915(c) waiver	165
	Elderly/Physically Disabled	Community Care Services Program	Medicaid 1915(c) waiver	98
Kentucky	Elderly/Physically Disabled	Home and Community Based Waiver	Medicaid 1915(c) waiver. ADC delivered through HCBS; not state funded.	150
	ID/DD	Supports for Community Living Waiver	Medicaid 1915(c) waiver	37
	TBI	Acquired Brain Injury Waiver	Medicaid 1915(c) waiver	26
Louisiana	Elderly/Physically Disabled	Adult Day Health Care Waiver	Medicaid 1915(c) waiver	112
	Elderly/Physically Disabled	Community Choices Waiver	Medicaid 1915(c) waiver	302

State	Population Category	HCBS Program	Funding Authority	Number of Total Returned Surveys
	Elderly/Physically Disabled	Long Term Personal Care Services Program	Medicaid State plan option	150
	ID/DD	New Opportunities Waiver	Medicaid 1915(c) waiver	146
Maryland	Elderly/Physically Disabled	Community Options Waiver	Medicaid1915(c) waiver	116
	TBI	Traumatic Brain Injury	Medicaid1915(c) waiver	0*
Minnesota	SMI	Personal Care Assistance Program	Medicaid State plan option	155
	Elderly	Elderly Waiver	Medicaid 1915(c) waiver	155
	TBI	Brain Injury Waiver	Medicaid 1915(c) waiver	72
New Hampshire	Elderly/Physically Disabled	Choices for Independence Home and Community Based Care Waiver	Medicaid 1915(c) waiver	147
	ID/DD	Developmental Disabilities Waiver	Medicaid 1915(c) waiver	91
	TBI	Acquired Brain Disorder Waiver	Medicaid 1915(c) waiver	20
	SMI	Bureau of Behavioral Health, Community Mental Health Services	Medicaid State plan, NH general funds, private insurance	174
Tennessee	Elderly/Physically Disabled	TennCare CHOICES in Long-Term Care	Medicaid 1115 waiver	304

*There are 0 completes because of a combined effect of a low number of individuals in the TBI program and the data collection ended before the vendor was able to begin data collection.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

There were 2,336 completed HCBS EoC surveys from 26 Medicaid HCBS programs included in the analysis of the survey data. The breakdown of individuals who completed the survey included:

- 70.2 percent in programs serving elderly (age 65+) Medicaid beneficiaries with disabilities, or programs serving working age (age 18-64) Medicaid beneficiaries with physical disabilities ;

- 8.3 percent served by programs for Medicaid beneficiaries with intellectual or developmental disabilities;
- 8.7 percent enrolled in programs targeting Medicaid beneficiaries with a traumatic brain injury; and
- 13.0 percent enrolled in Medicaid and receiving services due to a serious mental illness.

Demographics for those completing the survey included:

- Race: White 63.6%, Black 28.7%, Other Race 7.7%
- Language: English 90.8%, Spanish 3.6%, other 5.5%;
- Gender: Male 36.7%, Female 63.6%;
- Age: 18-24 2.0%, 25-34 5.7%, 35-44 8.6%, 45-54 17.8%, 55-64 25.3%, 65-74 21.2%, 75+ 19.3%;
- Living Arrangement: Lives alone 56.5%, Lives with others 43.6%;
- Metropolitan Statistical Area: Yes 76.5%, No 23.5%.

Other characteristics for those completing the survey included:

- Self-reported general health: Good, Very Good or Excellent 47.6%, Fair or Poor 52.4%
- Self-reported mental health: Good, Very Good or Excellent 68.3%, Fair or Poor 31.7%

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable. The same data were used for each aspect of testing below.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The respondent characteristics that were available and evaluated as potential case mix adjusters included self-reported general health rating, self-reported mental health rating, age , gender, and whether respondent lives alone. We also evaluated the differences in scores by HCBS population. Age, education, and health status are the most common CAHPS variables used in case mix adjustment. The Medicaid HCBS population, by definition, has low income; therefore, income was not used as a case mix adjuster. Education was initially considered as an item, but there were problems with face validity, namely some participants could not answer because of cognitive impairment due to developmental disability. Others had a college degree or higher but a traumatic brain injury left them cognitively less able than many high school students. Thus, the team opted not to include an education item. The survey was translated into Spanish, but the number of respondents responding in Spanish (46 respondents) were too few to conduct a comparison.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)
2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

We estimated Cronbach's Alpha values to assess internal consistency reliability, of survey items used in the scale measures. Cronbach's Alpha is a common measure for surveys with scale-type questions. A scale should have an alpha of 0.70 or greater to be considered reliable.[1]

We also looked at HCBS program-level reliability, or inter-unit reliability (IUR). Unit-level reliability indicates the extent to which the experiences of respondents within a unit (e.g., HCBS program) correlate with one another compared to the amount that reported experiences differ among units. As such, it reflects the signal-to-noise ratio; that is, the fraction of total variation due to signal (true variation in scores across units). One of the primary purposes of these measures is to be able to detect difference among HCBS programs, and thus, this ratio is a good indicator of the extent to which the scale measures and other survey items accomplish this goal. It also indicates how reliable a measure is across different respondents. This statistic represents a transformation of the F-statistic for testing differences among programs on a measure (IUR = (F-1)/F). IUR can be interpreted as the fraction of the variation among HCBS program scores that is due to real differences, rather than due to chance. If the IUR is higher, the ability of the item or scale measure to discriminate across programs is greater. Scales with reliability coefficients above 0.70 provide adequate precision for use in statistical analysis of unit-level comparisons.[2] As the IUR gets smaller, a larger sample is needed in order to reliably discriminate across programs.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Below are Cronbach's Alpha values for scale measures (Exhibit 2) and inter-unit reliability (IUR) statistics for all measures (Exhibit 3). Please reference tab 1.b.2a in the supplementary tables file for item-level IUR statistics for survey items used in the scale measures in Exhibit 3.

Exhibit 2. Cronbach's Alpha Values for Scale Measures

[1] Nunnally JC, Bernstein IH (1994). Psychometric Theory. New York: McGraw Hill.

[2] Nunnally, J. C. (1978). *Psychometric theory* (2nd ed). New York: McGraw-Hill .

Staff are reliable and helpful	0.84
Staff listen and communicate well	0.84
Case manager is helpful	0.82
Choosing the services that matter to you	0.50
Transportation to medical appointments	0.70
Personal safety and respect	0.17
Planning your time and activities	0.55

Exhibit 3. HCBS Inter-unit reliability (IUR) Statistics

Staff are reliable and helpful	0.66
Staff listen and communicate well	0.70
Case manager is helpful	0.38
Choosing the services that matter to you	0.77
Transportation to medical appointments	0.68
Personal safety and respect	0.32
Planning your time and activities	0.44
Overall Rating of Personal Assistance/Behavioral Health Staff	0.43
Would Recommend Personal Assistance/Behavioral Health to Family and Friends	0.55
Overall Rating of Homemaker	0.42
Would Recommend Homemaker to Family and Friends	0.76
Overall Rating of Case Manager	0.57
Would Recommend Case Manager to Family and Friends	0.48

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The Cronbach's Alpha scores range from 0.84 to 0.17, with three measures falling below the recommended 0.70 threshold. These were *Planning your time and activities (0.55), Choosing the services that matter to you* (0.50), and *Personal safety and respect* (0.17). While these values are below the recommended threshold, these measures were all deemed critical by the technical expert panel for assessing the quality of a HCBS program.

The IUR values range from 0.77 to 0.32, with the majority of measures (10/13) falling below the 0.70 threshold. This indicates that these measures will need a larger sample size to effectively discriminate

among programs. However, there are other important goals for using these measures, such as quality improvement for the states, where these measures will still be important.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Criterion validity refers to the extent to which the HCBS scale measures agree with some criterion of the "true" value of the measure, and can be predictive or concurrent. To evaluate the latter, we estimated correlation coefficients between each global rating measure and each scale measure. If the scale measures have good concurrent validity, then they should have a moderate to strong correlation ($r > 0.30$) with a conceptually related global rating measure. For example, we expect a strong correlation between the *Overall Rating of Case Manager* with the *Case Manager is Helpful* scale measure.

We also examined correlations among the scale measures to determine if they measure different constructs. As these are all measures of beneficiary experience with HCB services, we expect these factors to be related; however, all inter-scale measure correlations should be below 0.80 to indicate that these 7 factors, while related, do not overlap to the point of being redundant.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Exhibit 4. Correlation of Scale Measures and Related Global Rating Measures

Measure	**Correlation with Global Rating of Personal Assistance Staff**
Staff are reliable and helpful	0.36*
Staff listen and communicate well	0.37*
Personal safety and respect	0.24*
Measure	**Correlation with Global Rating of Homemaker**
Staff are reliable and helpful	0.29*
Staff listen and communicate well	0.33*
Personal safety and respect	0.19*
Measure	**Correlation with Global Rating of Case Manager**
Case manager is helpful	0.38*
Choosing the services that matter to you	0.33*

*p <.001

Exhibit 5. Inter-Scale Correlations

Final Scale Measures	Staff are reliable and helpful	Staff are reliable and helpful	Case manager is helpful	Choosing the services that matter to you	Transportation to medical appointments	Personal safety and respect	Planning your time and activities
Staff are reliable and helpful	1.00	-	-	-	-	-	-
Staff listen and communicate well	0.49	1.00	-	-	-	-	-
Case manager is helpful	0.24	0.21	1.00	-	-	-	-
Choosing the services that matter to you	0.12	0.12	0.11	1.00	-	-	-
Transportation to medical appointments	0.32	0.35	0.27	0.07	1.00	-	-
Personal safety and respect	0.22	0.32	0.28	0.11	0.23	1.00	-
Planning your time and activities	0.26	0.23	0.19	0.08	0.32	0.27	1.00

*All correlations are statistically significant at p <.001

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

For most measures, the correlations between the scale measures and the related global rating measures were moderate, suggesting that the scale measures are valid measures of beneficiary experience with these providers. The correlation for Personal Safety and Respect was low; however, it should be noted that there was not much variance in the items for this measure.

The scale measures were somewhat correlated with each other as they are all measures of beneficiary experience. However, no values were above 0.80, suggesting that these scales are measuring unique concepts.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without*

exclusion)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with user-selected risk factors***
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

*The CAHPS analysis program was employed as the statistical risk model, and this program allows researchers to select adjustment factors.

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The goals of case-mix adjustment are to help remove the effects of individual respondent characteristics that may affect ratings, remove effects that might be considered spurious (i.e., that reflect something other than quality of care), and remove incentives for providers to avoid "hard-to-treat" individuals. The most common CAHPS case-mix adjusters are age, education, and health status (both general health and emotional/mental health).

Three conditions were required in the selection of variables for case-mix adjustment:

- Within reporting units (HCBS programs), the case-mix variables must be related to the outcome measures (ratings). That is, the variables must have sufficient predictive power in relation to the outcomes (e.g., older respondents give higher ratings of their care). These variables are referred to as "predictors" of the outcome being examined.
- There must be variation between reporting units (HCBS programs) on these predictor variables. That is, the predictors must be unevenly distributed across reporting units (e.g., one program might have a population that tends to be much younger than the population of another program). This condition is the heterogeneity factor of the predictor.

- The case-mix variables must be appropriate for adjustment because they are not themselves determined by the provider's actions. That is, they must be characteristics that are brought to the program by the beneficiary (e.g., age or education), not characteristics that might be consequences of the beneficiary's satisfaction with, or assessment of, the program (e.g., number of visits with a provider). Predictors that are consequences of the beneficiary's satisfaction with the program are endogenous.

We tested the beneficiary characteristics of age, health status (both general health and emotional/mental health), gender, and whether the respondent lived alone as case-mix adjusters. These characteristics typically have the strongest and most consistent associations with patient-reported problems in other CAHPS surveys.[3] We also tested several survey design characteristics – survey mode (in-person vs. phone) and response option (standard vs. alternate[4]) -- as potential case mix adjusters.[5] The document "Instructions for Analyzing Data from CAHPS Surveys" dated April 2012 (available at: https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf) contains instructions for coding these variables and for including them in analyses using the CAHPS Analysis Program in SAS.

Our analysis for case-mix selection followed four steps:
1. Selection of potential case mix adjusters;
2. Estimation of heterogeneity;
3. Estimation of predictive power of the selected adjusters; and
4. Estimation of the impact of each adjuster.

The research team used stepwise regression to select a subset of the potential case-mix adjusters for further analysis. Stepwise regression analyses evaluated the strength of the relationship of each potential adjuster to ten global rating and scale measures in separate models in which each measure was regressed on all of the potential adjusters. In the stepwise regression models, the potential adjuster variables are added one by one to the model. For a variable to remain in the model, its F-statistic had to be significant at $p < 0.05$. Upon addition of a new variable to the model, each variable already in the model was reassessed, and variables that no longer retained an F-statistic significant at the retention p-level ($p < 0.05$) were excluded from the model. Only after this check was made and the necessary deletions accomplished was another variable added to the model. The stepwise process was complete

[3] O'Malley AJ, Zaslavsky AM, Elliott MN, Zaborski L, Cleary PD. (2005) Case-mix adjustment of the CAHPS Hospital Survey. *Health Serv Res.* Dec;40(6 Pt 2):2162-81.

[4] The research team opted to have two different response options for many of the survey items: the standard 4-point CAHPS frequency response (never, sometimes, usually, and always) and an alternate binary response (mostly yes and mostly no). This allows respondents who can use the 4-point frequency response to do so; for those that cannot, they are still able to participate in the survey using a modified response version. Similarly, based on input from the CAHPS Consortium and Julie Brown, the research team included the two different response scales for the global rating measures.

[5] Elliott MN, Zaslavsky AM, Goldstein E, Lehrman W, Hambarsoomians K, Beckett MK, Giordano L. (2009) Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. *Health Serv Res.* Apr;44(2 Pt 1):501-18. doi: 10.1111/j.1475-6773.2008.00914.x.

for a given model when none of the variables outside the model had an F statistic significant at $p < 0.05$ and every variable in the model was statistically significant at $p < 0.05$. Adjuster variables selected in any of the models formed a core set of potential case mix adjusters eligible for final selection.

The research team then estimated the **heterogeneity factor**, **predictive power**, **explanatory power**, and **impact factor** for each potential case-mix variable selected in the regression models. **Heterogeneity of the predictor variables** across programs was measured as the ratio of between-program to within-program variance of the residuals when the variable was regressed on all other potential case-mix adjusters in a random effects model, where the program was included in the model as a random effect. **Heterogeneity of outcome variables** across programs was measured as the ratio of between-program to within-program variance of the residuals when the variable was regressed on program in a random effects model. The research team measured **predictive power** as the incremental amount of variance explained by the predictor (represented as the partial r2 x 1,000) in the stepwise regression analyses, controlling for the other potential case-mix adjusters. To measure **explanatory power**, which considers both the predictive power of each potential adjuster and the heterogeneity of the adjusters across programs, the predictive power was multiplied by the adjuster heterogeneity factor. Finally, the research team calculated the **impact factor**, which standardizes explanatory power with respect to the overall variance in the outcome being assessed as explanatory power/outcome heterogeneity. Variables that had an impact factor >1.0 were considered as candidates for case mix adjusters.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Results are shown in Exhibits 6-8 below.

Exhibit 6. Parameter Estimates and Selection Status for Variable Selection Models - PCA, Homemaker and Case Manager Global Rating Measures

		Personal Assistance/Behavioral Health Staff Rating		Homemaker Rating		Case Manager Rating	
		Outcome Heterogeneity=0.004		Outcome Heterogeneity=0.013		Outcome Heterogeneity=0.023	
Case-mix Adjustment Variables	*Adjuster Heterogeneity*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*
Mental health	0.038	0.008	86.05	0.0114	33.49	0.010	17.31
Age (18-34)	0.161	0.0062	271.96	-	-	-	-
Age (25-34)	0.333	-	-	-	-	-	-
Age (35-44)	0.137	-	-	-	-	-	-
Age (45-54)	0.067	-	-	-	-	0.0024	7.091
Age (65-74)	0.191	-	-	-	-	0.0037	31.359
Age (75+)	0.306	-	-	-	-	0.0019	25.736
Survey mode	0.030	-	-	-	-	-	-
Response option mode	0.039	0.039	422.97	0.026	80.04	0.046	79.90
General health	0.041	-	-	0.005	15.40	-	-

Respondent lives alone	0.049	-	-	-	-	-	-
Gender	0.046	-	-	-	-	-	-

* Impact factor = (Adjuster Heterogeneity x (R^2 x 1,000)) / (Outcome heterogeneity)
Dashes indicate that the variable was not selected into the stepwise model

Exhibit 7. Parameter Estimates and Selection Status for Variable Selection Models – Getting Needed Care, Communication, and Case Management Scale Measures

		Getting Needed Care		Communication		Case Management	
		Outcome Heterogeneity= 0.044		Outcome Heterogeneity= 0.020		Outcome Heterogeneity= 0.008	
Case-mix Adjustment Variables	*Adjuster Heterogeneity*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*
Mental health	0.038	0.021	1.690	0.0035	6.624	0.0043	20.526
Age (18-34)	0.161	-	-	-	-	-	-
Age (25-34)	0.333	-	-	-	-	-	-
Age (35-44)	0.137	-	-	-	-	-	-
Age (45-54)	0.067	-	-	-	-	-	-
Age (65-74)	0.191	0.0039	16.780	-	-	-	-
Age (75+)	0.306	-	-	0.0022	33.875	-	-
Survey mode	0.030	-	-	0.0027	4.057	0.004	15.160
Response option mode	0.039	0.015	13.019	0.0106	21.022	-	-
General health	0.041	-	-	0.0018	3.720	-	-
Respondent lives alone	0.049	-	-	-	-	-	-
Gender	0.046	-	-	-	-	-	-

* Impact factor = (Adjuster Heterogeneity x (R^2 x 1,000)) / (Outcome heterogeneity)
Dashes indicate that the variable was not selected into the stepwise model

Exhibit 8. Parameter Estimates and Selection Status for Variable Selection Models – Choosing Your Services, Transportation, Personal Safety, and Community Inclusion Scale Measure Score Scale Measures

		Choosing Your Services		Transportation		Personal Safety		Community Inclusion	
		Outcome Heterogeneity= 0.033		Outcome Heterogeneity = 0.018		Outcome Heterogeneity= 0.003		Outcome Heterogeneity= 0.012	
Case-mix Adjustment Variables	*Adjuster Heterogeneity*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*	*Partial r^2*	*Impact Factor* >1.0*
Mental health	0.038	0.0139	15.677	0.0028	5.709	0.0094	104.046	0.0368	118.810
Age (18-34)	0.161	-	-	-	-	-	-	-	-
Age (25-34)	0.333	-	-	-	-	-	-	-	-
Age (35-44)	0.137	-	-	0.0038	28.231	-	-	-	-

Age (45-54)	0.067	-	-	-	-	-	-	-	-
Age (65-74)	0.191	-	-	-	-	-	-	-	-
Age (75+)	0.306	-	-	-	-	-	-	-	-
Survey mode	0.030	-	-	-	-	-	-	0.0043	11.022
Response option mode	0.039	-	-	0.0181	38.676	-	-	0.0266	89.995
General health	0.041	-	-	0.0085	18.927	-	-	0.0111	39.135
Respondent lives alone	0.049	-	-	0.0029	7.763	0.002	29.062	-	-
Sex	0.046	-	-	-	-	0.0033	44.913	-	-

* Impact factor = (Adjuster Heterogeneity x (R^2 x 1,000)) / (Outcome heterogeneity)
Dashes indicate that the variable was not selected into the stepwise model

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

See sections 1.8, 2b4.3. and 2b4.4a.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

See sections 1.8, 2b4.3. and 2b4.4a.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable.

2b4.9. Results of Risk Stratification Analysis:
Not applicable.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and*

what are the norms for the test conducted)

Variables that had an impact factor >1.0, and were therefore eligible to be considered as case- mix adjusters, included general health rating, mental health rating, age, gender, whether respondent lives alone, survey administration mode, and response option. Future administrations may also wish to include proxy status if assistance with the survey is permitted. Some CMS CAHPS surveys include adjustments for both proxy assisted and proxy completed questionnaires.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We used t-tests to compare the case-mix adjusted mean scores of each item, scale score, and global rating for each HCBS program within a state to the mean score of all programs combined within the state. A p-value of <0.05 was used to determine whether the scores were statistically significantly different from each other.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Exhibit 9 shows counts of programs that were statistically significantly different above or below their state mean for each measure. The exhibit also reports the percentage of programs that were statistically significant in either direction from their state mean.

Exhibit 9. Number and Percentage of Programs with Scores Differing from State Mean

Item #	Measure	Number of Programs Above Respective State Mean	Number of Programs Below Respective State Mean	% of Programs Differing from State Mean
Global Ratings Measures				
35	Global Rating of Personal Assistance/Behavioral Health Staff	6	2	32.0%
46	Global Rating of Homemaker	5	4	50.0%
54	Global Rating of Case Manager	5	4	36.0%

Item #	Measure	Number of Programs Above Respective State Mean	Number of Programs Below Respective State Mean	% of Programs Differing from State Mean
Recommendation Measures				
36	Recommendation of Personal Assistance/Behavioral Health Staff	6	3	36.0%
47	Recommendation of Homemaker	5	1	33.3%
55	Recommendation of Case Manager	1	1	8.0%
Scale Measures				
Staff are reliable and helpful		**6**	**2**	**33.3%**
13	Staff come to work on time	6	3	36.0%
14	Staff work as long as they are supposed to	5	3	32.0%
15	Someone tells you if staff cannot come	6	6	48.0%
19	Staff make sure you have enough privacy for dressing, showering, bathing	6	2	33.3%
37	Homemakers come to work on time	2	2	22.2%
38	Homemakers work as long as they are supposed to	4	2	33.3%
Staff listen and communicate well		**2**	**5**	**29.2%**
28	Staff are nice and polite	7	7	56.0%
29	Staff explanations are easy to understand	8	7	60.0%
30	Staff treat you the way you want them to	7	4	44.0%
31	Staff explain things in a way that is easy to understand	0	1	4.0%
32	Staff listen carefully to you	7	5	48.0%
33	Staff know what kind of help you need with everyday activities	4	2	24.0%
41	Homemakers are nice and polite	9	5	77.8%
42	Homemaker explanations are easy to understand	5	5	55.6%
43	Homemakers treat you the way you want them to	10	6	88.9%
44	Homemakers listen carefully	2	1	16.7%
45	Homemakers know what kind of help you need	1	2	16.7%
Case manager is helpful		**7**	**2**	**37.5%**
49	Able to contact this case manager when needed	9	5	56.0%
51	Case manager helped when asked for help with getting or fixing equipment	4	1	20.0%

Item #	Measure	Number of Programs Above Respective State Mean	Number of Programs Below Respective State Mean	% of Programs Differing from State Mean
53	Case manager helped when asked for help with getting other changes to services	4	2	24.0%
Choosing the services that matter to you		**8**	**5**	**54.2%**
56	Person-centered service plan included all of the things that are important	11	5	64.0%
57	Case manager knows what's on the service plan, including the things that are important	4	1	20.0%
Transportation to medical appointments		**7**	**7**	**58.3%**
59	Always have a way to get to your medical appointments	7	5	48.0%
61	Able to get in and out of this ride easily	7	6	52.0%
62	Ride arrives on time to pick you up	6	5	44.0%
Personal safety and respect		**3**	**4**	**29.2%**
64	Have someone to talk to if someone hurts you or does something to you that you don't like	3	1	16.0%
65	None of the staff take money or things without asking*	3	2	20.0%
68	None of the staff yell, swear, or curse*	5	3	32.0%
Planning your time and activities		**0**	**5**	**20.8%**
75	Can get together with nearby family	4	5	36.0%
77	Can get together with nearby friends	1	0	4.0%
78	Can do things in community	7	11	72.0%
79	Needs more help to do things in community	2	3	20.0%
80	Takes part in deciding what to do with their time	5	6	44.0%
81	Takes part in deciding when they do things each day	8	3	44.0%
Unmet Needs Measures				
18	There are no staff to help dress, shower, or bathe	0	0	0.0%
22	Sufficient staff to help you with meals	0	4	25.0%
25	Sufficient staff to help you with medications	0	0	0.0%
27	Sufficient staff to help you with toileting	6	3	40.9%
40	Sufficient homemakers to help you with household tasks	0	0	0.0%
Physical Safety Measure				

Item #	Measure	Number of Programs Above Respective State Mean	Number of Programs Below Respective State Mean	% of Programs Differing from State Mean
71	Do any staff hit or hurt you	5	0	20.0%

*Programs marked as above or below state means were statistically significantly different at p<.05

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The findings demonstrate that the measures produce results that adequately discriminate between service recipients' experience of care in their program compared to all programs within a state.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Not applicable.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

We conducted a nonresponse bias analysis to evaluate whether respondents and nonrespondents differed significantly. Response bias could be present if there is evidence that the responding population differed in important ways from the population of interest. Our response bias analysis involved comparing respondents to nonrespondents by mode of survey administration, HCBS population, and demographic characteristics using bivariate cross tabulations with chi-square tests (differences were considered statistically significant at $p < 0.05$).

The research team evaluated whether respondents and nonrespondents differed significantly across various characteristics using available data from the sample frame. Complete sample frame data were available only for a subset of the states; therefore, the total number of respondents for the nonresponse bias analysis is fewer than in the psychometric analyses.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Respondents and nonrespondents differed significantly by HCBS population, metropolitan statistical area (MSA) residence, state of residence, and guardian status. For example, more respondents were in the disabled (< age 65) group than non-respondents (42 percent vs. 36 percent, respectively); more respondents lived in an MSA than nonrespondents (77 percent vs. 74 percent, respectively); and more nonrespondents reported having a guardian than respondents (10 percent vs. 4 percent, respectively). There were no differences in response by assigned survey administration mode, survey response option, gender, or primary language.

Exhibit 10. Sample Frame Demographic Characteristics

Characteristics	Nonrespondents n=13,940	Respondents n=1,624	Total (Nonrespondents and Respondents Combined) N=15,564
HCBS Population*			
Aged (65+)	34.0	31.0	33.7
Disabled (<65)	36.4	41.8	36.9

Characteristics	Nonrespondents n=13,940	Respondents n=1,624	Total (Nonrespondents and Respondents Combined) N=15,564
ID/DD	19.0	11.3	18.2
TBI	4.2	6.3	4.4
SMI	6.4	9.6	6.8
Primary Language			
English	97.1	97.7	97.2
Spanish	2.0	1.9	2.0
Other	0.9	0.4	0.8
Metropolitan Statistical Area*			
Yes	74.3	76.5	74.5
No	25.7	23.5	25.5
Gender			
Male	41.9	43.0	42.0
Female	58.2	57.0	58.0
Assigned Survey Response			
Alternate	50.1	49.0	49.9
Standard CAHPS	50.0	51.1	50.1
Assigned Survey Mode			
In-person	80.6	79.2	80.4
Phone	19.4	20.8	19.6
State†*			
AZ	9.4	11.4	9.6
CO	17.7	15.0	17.4
GA	14.1	16.2	14.3
MD	19.2	7.1	18.0

Characteristics	Nonrespondents n=13,940	Respondents n=1,624	Total (Nonrespondents and Respondents Combined) N=15,564
MN	14.5	23.7	15.4
NH	25.2	26.6	25.3
Guardian*			
Yes	10.3	4.0	9.7
No	89.7	96.0	90.4

*Nonrespondents and respondents significantly differ by this characteristics at p <0.05

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Respondents and nonrespondents did differ by HCBS population, which will be a challenge with future data collection efforts. The team had difficulty reaching desired response rates from beneficiaries with intellectual or developmental disabilities. Future survey administrations may consider allowing proxy assistance with the survey, which will likely increase response rates.

In addition, to address this challenge, the following suggestions have been made to enhance future response rates. They include:

- Insuring that pre-notification letters originate from the state agency operating the HCBS program being surveyed so that those receiving the letter have familiarity with the letterhead.
- Ensuring that beneficiary contact information is accurate by requiring that the state and/or case managers verify beneficiary and guardian contact information for persons sampled.
- Ensuring that survey vendors have experience and specialized qualifications with the populations being surveyed so they are sensitized to particular considerations in interacting with people with certain types of disability. This is likely to increase rapport and result in improved recruitment.
- Targeting survey mode to persons/groups more likely to respond to a certain mode (rather than randomization to mode as happened in the field test).
- Conducting outreach to relevant stakeholders about the survey. This includes case managers and providers so they can encourage beneficiaries to participate when they receive inquiries from sampled members about the legitimacy of the survey. It may also include family and caregiver support groups. Stakeholders are more likely to encourage survey participation if they understand who is sponsoring the survey, its purpose and benefits.
- Not fielding the survey during the winter holiday season.
- Not fielding the survey during winter months in colder climates due to the risk of inclement weather prohibiting travel.

The lowest response rate was with the ID/DD population. Several of the field test study sites have also sponsored a

different, but similar survey – the National Core Indicators survey (NCI) -- that elicits feedback from people with ID/DD. Some state ID/DD agencies have conducted the NCI repeatedly over many years. Consequently beneficiaries, family members and guardians are very familiar with the survey.

Four field test states shared the response rates that they have attained in recent years for the adult NCI survey:

- KY: 94.5% response rate;
- AZ: 87% response rate
- CO: 39% response rate:
- CT: In order to accomplish target of 400 surveys, CT pulls a sample of upwards of 1,000.

In addition to the information provided by the TEFT states, the National Association for Directors of Developmental Disability Services (NASDDDS), one of the sponsors of the NCI survey for people with ID/DD, states that for their face-to-face surveys: *"Most states interview about 500 people to get the 400 sample size number (and most have to pull about 800 names to get the sample size)."* (http://www.nasddds.org/uploads/files/NCI_Description_and_Costs.pdf). While the NCI project does not report average response rates, this information from NASDDDS' website indicates the feasibility of achieving much higher response rates for the ID/DD subgroup than realized in the HCBS Experience of Care survey field test.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Other

If other: Collected by survey of beneficiaries

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

It is recommended that the HCBS EoC Survey be administered in-person or by phone. CATI or CAPI data collection is recommended which allow for the creation of electronic databases post data collection.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Data Collection:

- Despite a substantial amount of training and an extensive guide provided to survey vendors, all did not follow the data collection instructions exactly. These aspects can be reinforced when reviewing and modifying the materials for future administrations.
- In addition, implementers will need to be thoroughly educated about skip patterns in the EoC survey instrument, applicability of questions to their programs, and how to explain this to data collectors and survey programmers (who will need to take these patterns into effect when analyzing the data). Some of these skip patterns may be adapted to specific states, in which case additional work will be required with survey vendors (e.g., to explain why the skip patterns were adapted and conduct additional review of the field disks to ensure the surveys were appropriately adapted).
- It will be important for states to provide clear specifications about the nature of the work and realistic information about the context in which vendors will need to work. This is especially critical if they decide to use a survey vendor that is not familiar with the data collection instrument or HCBS populations.

Sampling:

- We recommend screening the sample for deceased individuals to the greatest extent possible.

Response Rate:

- Many beneficiaries of Medicaid HCBS programs have guardians from whom consent for the beneficiary's participation in a survey must be secured. For many states, this information is not centrally or readily available, or not updated. Accessing this information prior to contact will help increase participation.
- The AAPOR response rates considers individuals who are deceased or who are physically or mentally unable to respond as eligible respondents resulting in lower response rates. An alternative is to calculate a response rate that does not include such individuals as eligible respondents.
- To avoid alarming potential survey participants and to enhance the recruitment process, any pre-notification letters to the beneficiary should clearly identify the primary survey vendor.
- Programs should employ additional strategies for recruiting challenging populations, including using proxies. Additional outreach can involve case/care managers, or states might enlist advocacy groups to communicate to beneficiaries the importance of participating in the survey.

Timing of Data Collection:

- States that experience snow/ice during the winter should be encouraged to schedule data collection in other seasons.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

The final HCBS EoC survey will be available to state Medicaid Agencies for use free of charge. In addition to the survey instrument, users will have access to comprehensive materials supporting fielding, analysis, and reporting as well as CAHPS Analysis Program that performs analysis and significance testing.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

Testing of the survey and measures was recently completed. Plans for voluntary use by HCBS programs are underway.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The measures in this submission derive from the Home and Community-Based Services Experience of Care Survey. The survey was developed with CAHPS principles and the survey is currently under review by AHRQ and the CAHPS Consortium for a CAHPS trademark. Once the trademark has been received, the survey will be released publicly. Because the survey was developed for voluntary use in Medicaid HCBS programs, it is expected that many state Medicaid programs will begin using the survey within the next few years. Thus, it is expected that the measures derived from the survey will likely be used by states for their internal assessment of HCBS program quality and related quality improvement projects, as well as for public reporting at the state level. It is also possible that some measures may be considered as metrics in value based purchasing initiatives most typically associated with state Medicaid managed long-term services and supports. However, the survey and related measure use in state HCBS programs will be voluntary; at this time CMS has no plans to use the measures for national public reporting.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not applicable.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
See 4a.3.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** HCBS_EoC_NQF_Attachment.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare and Medicaid Services
Co.2 **Point of Contact:** Kerry, Lida, Kerry.Lida@cms.hhs.gov, 410-786-4826-
Co.3 **Measure Developer if different from Measure Steward:** Truven Health Analytics
Co.4 **Point of Contact:** Beth, Jackson, Beth.Jackson@truvenhealth.com, 508-520-1507-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
See section 1c.5. for list of technical expert panel members.
The research team involved in the development of the measures includes:
Centers for Medicare & Medicaid Services
Kerry Lida, Ph.D. kerry.lida@cms.hhs.gov
Michael R. Smith, MPA michael.smith1@cms.hhs.gov
Other Investigators
Beth Jackson, Ph.D., Truven Health Analytics
Susan Raetzman, M.S.P.H., Truven Health Analytics
Elizabeth Frentzel, M.P.H., American Institutes for Research

Coretta Mallery, Ph.D., American Institutes for Research Chris Pugliese, M.P.P., American Institutes for Research Lee Hargraves, Ph.D., American Institutes for Research Tandrea Hilliard, Ph.D., American Institutes for Research Chris Evensen, M.A., American Institutes for Research Steven Garfinkel, Ph.D., American Institutes for Research
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:** CMS is in the process of renewing their Measure Steward Agreement and received approval from NQF to submit these measures while this is in process.

Home and Community Based Services (HCBS) Experience of Care (EoC) Measures

Table of Contents

Home and Community Based Services (HCBS) Experience of Care (EoC) English Survey

Instructions for Vendor 56

Cognitive Screening Questions 58

Identification Questions 59

Getting Needed Services From Personal Assistant and Behavioral Health Staff 61

How Well Personal Assistant and Behavioral Health Staff Communicate and Treat You 65

Getting Needed Services From Homemakers 68

How Well Homemakers Communicate and Treat You 70

Your Case Manager 73

Choosing Your Services 75

Transportation 76

Personal Safety 77

Community Inclusion and Empowerment 80

About You 82

Interviewer Questions 86

Supplemental Employment Module 87

Home and Community Based Services (HCBS) Experience of Care (EoC) Spanish Survey

Instructions for Vendor 95

COGNITIVE SCREENING QUESTIONS 96

PREGUNTAS DE IDENTIFICACIÓN 97

Obtención de los servicios necesarios de parte de los auxiliares de cuidados personales y del personal de salud mental 99

Qué tan bien se comunica(n) con usted los auxiliares de cuidados personales o el personal de salud mental y qué tan bien lo(a) tratan 102

Obtención de los servicios necesarios de los ayudantes de oficios domésticos 105
Qué tan bien se comunican con usted los ayudantes de oficios domésticos y qué tan bien lo(a) tratan 106
Su encargado de caso 108
La elección de sus servicios..... 111
Transporte..... 111
Seguridad personal 112
Comunidad y empoderamiento..... 114
Sobre usted..... 116
Interviewer Questions 119
MÓDULO COMPLEMENTARIO SOBRE EMPLEO 120

Home and Community Based Services Experience of Care Survey

Version: 1.0

Population: Adult

Language: English

Response Scale: 4 point and 2 point alternative

Notes

- **Supplemental items:** Survey users may add questions to this survey. The supplemental items are available at the end of this survey.

Instructions for Vendor

- The interview is intended as an interviewer-administered survey, thus all text that appears in initial uppercase and lowercase letters should be read aloud. Text that appears in **bold, lowercase letters** should be emphasized.
- Text in {*italics and in braces*} will be provided by the HCBS program's administrative data. However, if the interviewee provides another term, that term should be used in place of the program-specific term wherever indicated. For example, some interviewees may refer to their case manager by another title, which should be used instead throughout the survey.
- For response options of "never," "sometimes," "usually," and "always," if the respondent cannot use that scale, the alternate version of the survey should be used which uses the response options of "mostly yes" and "mostly no." These response options are reserved for respondents who find the "never," "sometimes," "usually," "always" response scale cognitively challenging.
- For response options of 0 to 10, if the respondent cannot use that scale, the alternate version o f the survey should be used which uses the response options of "Excellent," "very good," "good," "fair," or "poor." These response options are reserved for respondents who find the numeric scale cognitively challenging.
- All questions include a "REFUSED" response option. In this case, "refused" means the respondent did not provide any answer to the question.
- All questions include a "DON'T KNOW" response option. This is used when the respondent indicates that he or she does not know the answer and cannot provide a response to the question.
- All questions include an "UNCLEAR" response option. This should be used when a respondent answers, but the interviewer cannot clarify the meaning of the response even after minor probing **or** the response is completely unrelated to the question—for example, the response to "Do your homemakers listen carefully to what you say?" is "I like to sit by Mary."
- Some responses have skip patterns, which are expressed as "→ GO TO Q#." The interviewer will be automatically skipped to the next correct item.
- Not all respondents have all services. Items Q4 through Q12 help to confirm which services a respondent has. The table after it presents the logic of which items should be used.
- **Use Singular/Plural as needed:** Modify items such that the interviewer can use the correct form (singular or plural) of the survey item.
- **Use Program-Specific Terms:** Where appropriate, add in the program-specific terms for staff (e.g., [*program-specific term for these types of staff*]) but allow the interviewer to modify the term based on the respondent's choice of the word. It will be necessary to obtain information for program-specific terms. State administrative data should include the following information:
  - Agency name(s)
  - Titles of staff who provide care
  - Names of staff who provide care
  - Activities that each staff member provides (this will help with identifying appropriate skip logic)

- Hours of staff who come to the home

Cognitive Screening Questions

People might be paid to help you get ready in the morning, with housework, go places, or get mental health services. This survey is about the people who are paid to help you in your home and community with everyday activities. It also asks about the services you get.

1. Does someone come into your home to help you?

[1] ☐ YES
[2] ☐ NO → END SURVEY
[-1] ☐ DON'T KNOW → END SURVEY
[-2] ☐ REFUSED → END SURVEY
[-3] ☐ UNCLEAR RESPONSE→ END SURVEY

2. How do they help you?

_____

[EXAMPLES OF CORRECT RESPONSES INCLUDE]

- HELPS ME GET READY EVERY DAY
- CLEANS MY HOME
- WORKS WITH ME AT MY JOB
- HELPS ME TO DO THINGS
- DRIVES ME AROUND

[-1] ☐ DON'T KNOW → GO TO THE ABOUT YOU SECTION
[-2] ☐ REFUSED → GO TO THE ABOUT YOU SECTION
[-3] ☐ UNCLEAR RESPONSE → GO TO THE ABOUT YOU SECTION

3. What do you call them?

_____

[EXAMPLES OF SUFFICIENT RESPONSES INCLUDE]

- MY WORKER
- MY ASSISTANT
- NAMES OF STAFF (JO, DAWN, ETC.)

[-1] ☐ DON'T KNOW → GO TO THE ABOUT YOU SECTION
[-2] ☐ REFUSED → GO TO THE ABOUT YOU SECTION
[-3] ☐ UNCLEAR RESPONSE → GO TO THE ABOUT YOU SECTION

CSQPASS.

(INT: IF ALL 3 QUESTIONS ANSWERED CORRECTLY, ENTER 1 TO CONTINUE.)

1 PASS - ALL 3 QUESTIONS WERE ANSWERED CORRECTLY → GO TO Q4

2 FAIL - AT LEAST 1 QUESTION WAS NOT ANSWERED CORRECTLY → GO TO SURVEND

SURVEND.

Thank you for your time. Those are all the questions we have.

Have a nice day/evening. (INT: ENTER 1 TO EXIT SURVEY)

Identification Questions

Now I would like to ask you some more questions about the types of people who come to your home.

4. Do you get {*program specific term for personal assistance*} at home?

- [1] ☐ YES
- [2] ☐ NO → GO TO Q6
- [-1] ☐ DON'T KNOW → GO TO Q6
- [-2] ☐ REFUSED → GO TO Q6
- [-3] ☐ UNCLEAR RESPONSE → GO TO Q6

5. What do you call the person or people who give you {*program-specific term for personal assistance*}? For example, do you call them {*program-specific term for personal assistance*}, staff, personal care attendants, PCAs, workers, or something else?

_____

[ADD RESPONSE WHEREVER IT SAYS "*personal assistance/behavioral health staff"*]

6. Do you get {*program specific term for behavioral health specialist services*} at home?

- [1] ☐ YES
- [2] ☐ NO → GO TO Q8
- [-1] ☐ DON'T KNOW → GO TO Q8
- [-2] ☐ REFUSED → GO TO Q8
- [-3] ☐ UNCLEAR RESPONSE → GO TO Q8

7. What do you call the person or people who give you {*program specific term for behavioral health specialist services*}? For example, do you call them {*program-specific term for behavioral health specialists*}, counselors, peer supports, recovery assistants, or something else?

_____

[ADD RESPONSE WHEREVER IT SAYS "*personal assistance/behavioral health staff" ;* IF Q4 IS ALSO= YES, LIST BOTH TITLES]

8. Do you get {*program specific term for homemaker services*} at home?

[1] ☐ YES
[2] ☐ NO → GO TO Q11
[-1] ☐ DON'T KNOW → GO TO Q11
[-2] ☐ REFUSED → GO TO Q11
[-3] ☐ UNCLEAR RESPONSE → GO TO Q11

9. What do you call the person or people who give you {*program specific term for homemaker services*}? For example, do you call them {*program-specific term for homemaker*}, aides, homemakers, chore workers, or something else?

_____

[ADD RESPONSE WHEREVER IT SAYS "*homemaker"*]

10. [IF (Q4 *OR* Q6) *AND* Q8= YES, ASK] Do the same people who help you with everyday activities also help you to clean your home?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

11. Do you get help from {*program specific term for case manager services}* to help make sure that you have all the services you need?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

12. What do you call the person who gives you {*program specific term for case manager services*}? For example, do you call the person a {*program-specific term for case manager*}, case manager, care manager, service coordinator, supports coordinator, social worker, or something else?

_____

[ADD RESPONSE WHEREVER IT SAYS "*case manager"*]

BELOW ARE INSTRUCTIONS TO WHICH QUESTIONS TO ASK FOR EACH RESPONSE ABOVE

ITEM AND RESPONSE	ACTION
IF Q4 OR Q6= YES, AND Q8 = NO, DON'T KNOW, REFUSE, UNCLEAR	ASK Q13-Q36, AND Q48 ONWARD
IF Q4 AND Q6 = NO	SKIP Q13-36, 57 AND 79
IF Q8 = YES	ASK Q37-Q47, AND Q48 ONWARD
IF Q10 = YES	ASK Q13-Q36, Q39, Q40, AND Q48 ONWARD
IF Q11 = ANY RESPONSE	ASK Q48 – Q55, AND Q56 ONWARD

Getting Needed Services From Personal Assistant and Behavioral Health Staff

13. First I would like to talk about the {*personal assistance/behavioral health staff*}who are paid to help you with everyday activities—for example, getting dressed, using the bathroom, taking a bath or shower, or going places. How often do {*personal assistance/behavioral health staff*} come to work on time? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: First I would like to talk about the {*personal assistance/behavioral health staff*}who are paid to help you with everyday activities—for example, getting dressed, using the bathroom, taking a bath or shower, or going places. Do {*personal assistance/behavioral health staff*} come to work on time? Would you say. . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

14. How often do {personal assistance/behavioral health staff} work as long as they are supposed to? Would you say...

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*personal assistance/behavioral health staff*} work as long as they are supposed to? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

15. Sometimes staff cannot come to work on a day that they are scheduled. When staff cannot come to work on a day that they are scheduled, does someone let you know if {*personal assistance/behavioral health staff*} cannot come that day?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

16. Do you need help from {*personal assistance/behavioral health staff*} to get dressed, take a shower, or bathe?

[1] ☐ YES
[2] ☐ NO → GO TO Q20
[-1] ☐ DON'T KNOW → GO TO Q20
[-2] ☐ REFUSED →GO TO Q20
[-3] ☐ UNCLEAR RESPONSE →GO TO Q20

17. Do you **always** get dressed, take a shower, or bathe when you need to?

[1] ☐ YES → GO TO Q19
[2] ☐ NO
[-1] ☐ DON'T KNOW → GO TO Q19
[-2] ☐ REFUSED → GO TO Q19
[-3] ☐ UNCLEAR RESPONSE → GO TO Q19

18. Is this because there are no {*personal assistance/behavioral health staff*} to help you?

[1] ☐ YES

2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

19. How often do {*personal assistance/behavioral health staff*} make sure you have enough personal privacy when you dress, take a shower, or bathe? Would you say. . .

1☐ Never,
2☐ Sometimes,
-3☐ Usually, or
-3☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*personal assistance/behavioral health staff*} make sure you have enough personal privacy when you dress, take a shower, or bathe? Would you say. . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

20. Do you need help from {*personal assistance/behavioral health staff*} with your meals, such as help making or cooking meals or help eating?

1☐ YES
2☐ NO → GO TO Q23
-1☐ DON'T KNOW → GO TO Q23
-2☐ REFUSED → GO TO Q23
-3☐ UNCLEAR RESPONSE → GO TO Q23

21. Are you **always** able to get something to eat when you are hungry?

1☐ YES → GO TO Q23
2☐ NO
-1☐ DON'T KNOW → GO TO Q23
-2☐ REFUSED → GO TO Q23
-3☐ UNCLEAR RESPONSE → GO TO Q23

22. Is this because there are no {*personal assistance/behavioral health staff*} to help you?

1☐ YES
2☐ NO
-1☐ DON'T KNOW

-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

23. Sometimes people need help taking their medicines, such as reminders to take a medicine, help pouring them, or setting up their pills. Do you need help from {*personal assistance/behavioral health staff*} to take your medicines?

1 ☐ YES
2 ☐ NO → GO TO Q26
-1 ☐ DON'T KNOW → GO TO Q26
-2 ☐ REFUSED → GO TO Q26
-3 ☐ UNCLEAR RESPONSE → GO TO Q26

24. Do you **always** take your medicine when you are supposed to?

1 ☐ YES → GO TO Q26
2 ☐ NO
-1 ☐ DON'T KNOW → GO TO Q26
-2 ☐ REFUSED → GO TO Q26
-3 ☐ UNCLEAR RESPONSE → GO TO Q26

25. Is this because there are no {*personal assistance/behavioral health staff*} to help you?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

26. Help with toileting includes helping someone get on and off the toilet or helping to change disposable briefs or pads. Do you need help from {*personal assistance/behavioral health staff*} with toileting?

1 ☐ YES
2 ☐ NO → GO TO Q28
-1 ☐ DON'T KNOW → GO TO Q28
-2 ☐ REFUSED → GO TO Q28
-3 ☐ UNCLEAR RESPONSE → GO TO Q28

27. Do you get all the help you need with toileting from {personal assistance/behavioral health staff} when you need it?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED

-3☐ UNCLEAR RESPONSE

How Well Personal Assistant and Behavioral Health Staff Communicate and Treat You

The next several questions ask about how {personal assistance/behavioral health staff} treat you.

28. How often are {personal assistance/behavioral health staff} nice and polite to you? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Are {*personal assistance/behavioral health staff*} nice and polite to you? Would you say . . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

29. How often are the explanations {*personal assistance/behavioral health staff*} give you hard to understand because of an accent or the way {*personal assistance/behavioral health staff}* speak English? Would you say …

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Are the explanations {*personal assistance/behavioral health staff*} give you hard to understand because of an accent or the way {*personal assistance/behavioral health staff*} speak English? Would you say. . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED

-3☐ UNCLEAR RESPONSE

30. How often do {*personal assistance/behavioral health staff*} treat you the way you want them to? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*personal assistance/behavioral health staff*} treat you the way you want them to? Would you say . . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

31. How often do {*personal assistance/behavioral health staff*} explain things in a way that is easy to understand? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*personal assistance/behavioral health staff*} explain things in a way that is easy to understand? Would you say . . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

32. How often do {*personal assistance/behavioral health staff*} listen carefully to you? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?

-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*personal assistance/behavioral health staff*} listen carefully to you?
Would you say . . .
1 ☐ Mostly yes, or,
2 ☐ Mostly no?
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

33. Do you feel {*personal assistance/behavioral health staff*} know what kind of help **you** need with everyday activities, like getting ready in the morning, getting groceries, or going places in your community?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

34. Do *{personal assistance/behavioral health staff}* encourage you to do things for yourself if you can?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

35. Using any number from 0 to 10, where 0 is the worst help from {*personal assistance/behavioral health staff*} possible and 10 is the best help from {*personal assistance/behavioral health staff*} possible, what number would you use to rate the help you get from {*personal assistance/behavioral health staff*}?

__0 TO 10
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: How would you rate the help you get from {*personal assistance/behavioral health staff*}? Would you say . . .
1 ☐ Excellent,
2 ☐ Very good,
3 ☐ Good,

4☐ Fair, or
5☐ Poor?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

36. Would you recommend the {*personal assistance/behavioral health staff}* who help you to your family and friends if they needed help with everyday activities? Would you say you recommend the {*personal assistance/behavioral health staff}* . . .

1☐ Definitely no,
2☐ Probably no,
3☐ Probably yes, or
4☐ Definitely yes?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

Getting Needed Services From Homemakers

The next several questions are about the {*homemakers*}, the staff who are paid to help you do tasks around the home—such as cleaning, grocery shopping, or doing laundry.

37. How often do {*homemakers*} come to work on time? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*homemakers*} come to work on time? Would you say . . .
1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

38. How often do {*homemakers*} work as long as they are supposed to? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*homemakers*} work as long as they are supposed to? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

39. Do your household tasks, like cleaning and laundry, **always** get done when you need them to? [ASK IF HOMEMAKER IS THE SAME AS PCA STAFF]

[1] ☐ YES →GO TO Q41
[2] ☐ NO
[-1] ☐ DON'T KNOW →GO TO Q41
[-2] ☐ REFUSED →GO TO Q41
[-3] ☐ UNCLEAR RESPONSE →GO TO Q41

40. Is this because there are no {*homemakers*} to help you? [ASK IF HOMEMAKER IS THE SAME AS PCA STAFF]

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

How Well Homemakers Communicate and Treat You

The next several questions ask about how {homemakers} treat you.

41. How often are {*homemakers*} nice and polite to you? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Are {*homemakers*} nice and polite to you? Would you say ...
[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

42. How often are the explanations {*homemaker*} give you hard to understand because of an accent or the way the {*homemakers}* speak English? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Are the explanations {*homemakers*} give you hard to understand because of an accent or the way {*homemakers*} speak English? Would you say. . .
[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

43. How often do {*homemakers*} treat you the way you want them to? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*homemakers*} treat you the way you want them to? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

44. How often do {*homemakers*} listen carefully to you? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Do {*homemakers*} listen carefully to you? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

45. Do you feel {*homemakers*} know what kind of help you need?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

46. Using any number from 0 to 10, where 0 is the worst help from {*homemakers*} possible and 10 is the best help from {*homemakers*} possible, what number would you use to rate the help you get from {*homemakers*}?

__0 TO 10
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: How would you rate the help you get from {*homemakers*}? Would you say . . .
1 ☐ Excellent,
2 ☐ Very good,
3 ☐ Good,
4 ☐ Fair, or
5 ☐ Poor?
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

47. Would you recommend the {*homemakers}* who help you to your family and friends if they needed {*program-specific term for homemaker services*}? Would you say you recommend the {*homemakers*} . . .

1 ☐ Definitely no,
2 ☐ Probably no,
3 ☐ Probably yes, or
4 ☐ Definitely yes?
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

Your Case Manager

Now I would like to talk to you about your {*case manager*}, the person who helps make sure you have the services you need.

48. Do you know who your {*case manager*} is?

1☐ YES
2☐ NO → GO TO Q56
-1☐ DON'T KNOW →GO TO Q56
-2☐ REFUSED →GO TO Q56
-3☐ UNCLEAR RESPONSE →GO TO Q56

49. Can you contact this {*case manager*} when you need to?

1☐ YES
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

50. Some people need to get equipment to help them, like wheelchairs or walkers, and other people need their equipment replaced or fixed. Have you asked this {*case manager*} for help with getting or fixing equipment?

1☐ YES
2☐ NO → GO TO Q52
3☐ DON'T NEED → GO TO Q52
-1☐ DON'T KNOW → GO TO Q52
-2☐ REFUSED → GO TO Q52
-3☐ UNCLEAR RESPONSE → GO TO Q52

51. Did this {*case manager*} work with you when you asked for help with getting or fixing equipment?

1☐ YES
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

52. Have you asked this {*case manager*} for help in getting any changes to your services, such as more help from {*personal assistance/behavioral health staff and/or homemakers if applicable*}, or for help with getting places or finding a job?

[1] ☐ YES
[2] ☐ NO → GO TO 54
[3] ☐ DON'T NEED → GO TO Q54
[-1] ☐ DON'T KNOW→ GO TO Q54
[-2] ☐ REFUSED → GO TO Q54
[-3] ☐ UNCLEAR RESPONSE → GO TO Q54

53. Did this {*case manager*} work with you when you asked for help with getting other changes to your services?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

54. Using any number from 0 to 10, where 0 is the worst help from {*case manager*} possible and 10 is the best help from {*case manager*} possible, what number would you use to rate the help you get from {*case manager*}?

__0 TO 10
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: How would you rate the help you get from the {*case manager*}? Would you say . . .
[1] ☐ Excellent,
[2] ☐ Very good,
[3] ☐ Good,
[4] ☐ Fair, or
[5] ☐ Poor?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

55. Would you recommend the {*case manager*} who helps you to your family and friends if they needed {*program-specific term for case-management services*}? Would you say you recommend the {*case manager*} . . .

1☐ Definitely no,
2☐ Probably no,
3☐ Probably yes, or
4☐ Definitely yes?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

Choosing Your Services

56. Does your [*program-specific term for "*service *plan"*] include . . .

1☐ **None** of the things that are important to you,
2☐ **Some** of the things that are important to you,
3☐ **Most** of the things that are important to you, or
4☐ **All** of the things that are important to you?
-1☐ DON'T KNOW → GO TO Q58
-2☐ REFUSED → GO TO Q58
-3☐ UNCLEAR RESPONSE → GO TO Q58

57. Do you feel {*personal assistance/behavioral health staff*} know what's on your [*program-specific term for "service plan"*], including the things that are important to you?

1☐ YES
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

58. Who would you talk to if you wanted to change your [*program-specific term for "service plan"*]? Anyone else? [INTERVIEWER MARKS ALL THAT APPLY]

1☐ CASE MANAGER
2☐ OTHER STAFF
3☐ FAMILY/FRIENDS
4☐ SOMEONE ELSE, PLEASE SPECIFY _____
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

Transportation

The next questions ask about how you get to places in your community.

59. Medical appointments include seeing a doctor, a dentist, a therapist, or someone else who takes care of your health. How often do you have a way to get to your medical appointments? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Medical appointments include seeing a doctor, a dentist, a therapist, or someone else who takes care of your health. Do you have a way to get to your medical appointments? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

60. Do you use a van or some other transportation service? Do not include a van you own.

[1] ☐ YES
[2] ☐ NO → GO TO Q63
[-1] ☐ DON'T KNOW → GO TO Q63
[-2] ☐ REFUSED → GO TO Q63
[-3] ☐ UNCLEAR RESPONSE → GO TO Q63

61. Are you able to get in and out of this ride easily?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

62. How often does this ride arrive on time to pick you up? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Does this ride arrive on time to pick you up? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

Personal Safety

The next few questions ask about your personal safety**.**

63. Who would you contact in case of an emergency? [INTERVIEWER MARKS ALL THAT APPLY]

[1] ☐ FAMILY MEMBER OR FRIEND
[2] ☐ CASE MANAGER
[3] ☐ AGENCY THAT PROVIDES HOME- AND COMMUNITY-BASED SERVICES
[4] ☐ PAID EMERGENCY RESPONSE SERVICE (E.G., LIFELINE)
[5] ☐ 9–1–1 (FIRST RESPONDERS, POLICE, LAW ENFORCEMENT)
[6] ☐ SOMEONE ELSE, PLEASE SPECIFY _____
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

64. Is there a person you can talk to if someone hurts you or does something to you that you don't like?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

The next few questions ask if anyone paid to help you now is treating you badly. This includes {*personal assistance/behavioral health staff, homemakers, or your case manager*}. We are asking everyone the next questions—not just you. [ADD STATE-SPECIFIC LANGUAGE HERE REGARDING MANDATED REPORTING, IF APPROPRIATE—"I want to remind you that, although your answers are confidential, I have a legal responsibility to tell {*STATE*} if I hear something that makes me think you are being hurt or are in danger."]

65. Do **any** of the {*personal assistance/behavioral health staff, homemakers, or your case managers*} that you have **now** take your money or your things without asking you first?

1☐ YES
2☐ NO → GO TO Q68
-1☐ DON'T KNOW → GO TO Q68
-2☐ REFUSED → GO TO Q68
-3☐ UNCLEAR RESPONSE → GO TO Q68

66. Is someone working with you to fix this problem?

1☐ YES
2☐ NO → GO TO Q68
-1☐ DON'T KNOW → GO TO Q68
-2☐ REFUSED → GO TO Q68
-3☐ UNCLEAR RESPONSE → GO TO Q68

67. Who is working with you to fix this problem? Anyone else? [INTERVIEWER MARKS ALL THAT APPLY

1☐ FAMILY MEMBER OR FRIEND
2☐ CASE MANAGER
3☐ AGENCY
4☐ SOMEONE ELSE, PLEASE SPECIFY _____
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

68. Do any {*staff*} that you have now yell, swear, or curse at you?

1☐ YES
2☐ NO → GO TO Q71
-1☐ DON'T KNOW → GO TO Q71
-2☐ REFUSED → GO TO Q71
-3☐ UNCLEAR RESPONSE → GO TO Q71

69. Is someone working with you to fix this problem?

1☐ YES
2☐ NO → GO TO Q71
-1☐ DON'T KNOW → GO TO Q71
-2☐ REFUSED → GO TO Q71
-3☐ UNCLEAR RESPONSE → GO TO Q71

70. Who is working with you to fix this problem? Anyone else? [INTERVIEWER MARKS ALL THAT APPLY]

1☐ FAMILY MEMBER OR FRIEND
2☐ CASE MANAGER
3☐ AGENCY
4☐ SOMEONE ELSE, PLEASE SPECIFY _____
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

71. Do any {*staff*} that you have now hit you or hurt you?

1☐ YES
2☐ NO → GO TO Q74
-1☐ DON'T KNOW → GO TO Q74
-2☐ REFUSED → GO TO Q74
-3☐ UNCLEAR RESPONSE → GO TO Q74

72. Is someone working with you to fix this problem?

1☐ YES
2☐ NO → GO TO Q74
-1☐ DON'T KNOW → GO TO Q74
-2☐ REFUSED → GO TO Q74
-3☐ UNCLEAR RESPONSE → GO TO Q74

73. Who is working with you to fix this problem? Anyone else? [INTERVIEWER MARKS ALL THAT APPLY]

1☐ FAMILY MEMBER OR FRIEND
2☐ CASE MANAGER
3☐ AGENCY
4☐ SOMEONE ELSE, PLEASE SPECIFY _____
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

Community Inclusion and Empowerment

Now I'd like to ask you about the things you do in your community.

74. Do you have any **family** members who live nearby? Do not include family members you live with.

[1] ☐ YES
[2] ☐ NO → GO TO Q76
[-1] ☐ DON'T KNOW → GO TO Q76
[-2] ☐ REFUSED → GO TO Q76
[-3] ☐ UNCLEAR RESPONSE → GO TO Q76

75. When you want to, how often can you get together with these family members who live nearby? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: When you want to, can you get together with these family members who live nearby? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

76. Do you have any **friends** who live nearby?

[1] ☐ YES
[2] ☐ NO → GO TO Q78
[-1] ☐ DON'T KNOW → GO TO Q78
[-2] ☐ REFUSED → GO TO Q78
[-3] ☐ UNCLEAR RESPONSE → GO TO Q78

77. When you want to, how often can you get together with these friends who live nearby? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: When you want to, can you get together with these friends who live nearby? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

78. When you want to, how often can you do things in the community that you like?

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: When you want to, can you do things in the community that you like? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

79. Do you need more help than you get now from {*personal assistance/behavioral health staff*} to do things in your community?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

80. Do you take part in deciding **what** you do with your time each day?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

81. Do you take part in deciding **when** you do things each day—for example, deciding when you get up, eat, or go to bed?

1 ☐ YES
2 ☐ NO
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

About You

Now I just have a few more questions about you.

82. In general, how would you rate your overall health? Would you say . . .

1 ☐ Excellent,
2 ☐ Very good,
3 ☐ Good,
4 ☐ Fair, or
5 ☐ Poor?
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

83. In general, how would you rate your overall mental or emotional health? Would you say . . .

1 ☐ Excellent,
2 ☐ Very good,
3 ☐ Good,
4 ☐ Fair, or
5 ☐ Poor?
-1 ☐ DON'T KNOW
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE

84. What is your age?

[1] ☐ 18 TO 24 YEARS → GO TO Q85
[2] ☐ 25 TO 34 YEARS → GO TO Q85
[3] ☐ 35 TO 44 YEARS → GO TO Q85
[4] ☐ 45 TO 54 YEARS → GO TO Q85
[5] ☐ 55 TO 64 YEARS → GO TO Q85
[6] ☐ 65 TO 74 YEARS → GO TO Q85
[7] ☐ 75 YEARS OR OLDER → GO TO Q85
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED→ GO TO Q85
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: In what year were you born?
_____ (YEAR)
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

85. [IF NECESSARY, ASK, AND VERIFY IF OVER THE PHONE] Are you male or female?

[1] ☐ MALE
[2] ☐ FEMALE
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

86. Are you of Hispanic, Latino, or Spanish origin?

[1] ☐ YES, HISPANIC, LATINO, OR SPANISH
[2] ☐ NO, NOT HISPANIC, LATINO, OR SPANISH → GO TO Q88
[-1] ☐ DON'T KNOW → GO TO Q88
[-2] ☐ REFUSED → GO TO Q88
[-3] ☐ UNCLEAR RESPONSE → GO TO Q88

87. Which group best describes you? [READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.]

[1] ☐ Mexican, Mexican American, Chicano, Chicana
[2] ☐ Puerto Rican
[3] ☐ Cuban
[4] ☐ Another Hispanic, Latino, or Spanish origin
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

88. What is your race? You may choose one or more of the following. Would you say you are…

[1] ☐ White → GO TO Q91
[2] ☐ Black or African-American → GO TO Q91
[3] ☐ Asian → GO TO Q89
[4] ☐ Native Hawaiian or other Pacific Islander → GO TO Q90
[5] ☐ American Indian or Alaska Native → GO TO Q91
[6] ☐ OTHER → GO TO Q91
[-1] ☐ DON'T KNOW → GO TO Q91
[-2] ☐ REFUSED → GO TO Q91
[-3] ☐ UNCLEAR RESPONSE → GO TO Q91

89. Which group best describes you? [READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.]

[1] ☐ Asian Indian → GO TO Q91
[2] ☐ Chinese → GO TO Q91
[3] ☐ Filipino → GO TO Q91
[4] ☐ Japanese → GO TO Q91
[5] ☐ Korean → GO TO Q91
[6] ☐ Vietnamese → GO TO Q91
[7] ☐ Other Asian → GO TO Q91
[-1] ☐ DON'T KNOW → GO TO Q91
[-2] ☐ REFUSED → GO TO Q91
[-3] ☐ UNCLEAR RESPONSE → GO TO Q91

90. Which group best describes you? [READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.]

[1] ☐ Native Hawaiian → GO TO Q91
[2] ☐ Guamanian or Chamorro → GO TO Q91
[3] ☐ Samoan → GO TO Q91
[4] ☐ Other Pacific Islander → GO TO Q91
[-1] ☐ DON'T KNOW → GO TO Q91
[-2] ☐ REFUSED → GO TO Q91
[-3] ☐ UNCLEAR RESPONSE → GO TO Q91

91. Do you speak a language other than English at home? [READ CHOICES ONLY IF NEEDED…]

[1] ☐ Yes
[2] ☐ No → GO TO Q93
[-1] ☐ DON'T KNOW → GO TO Q93
[-2] ☐ REFUSED → GO TO Q93
[-3] ☐ UNCLEAR RESPONSE → GO TO Q93

92. What is the language you speak at home?

[1] ☐ Spanish,
[2] ☐ Some other language → Which one? _____
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

93. [IF NECESSARY, ASK] How many adults live at your home, including you?

[1] ☐ 1 [JUST THE RESPONDENT] → END SURVEY
[2] ☐ 2 TO 3
[3] ☐ 4 OR MORE
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

94. [IF NECESSARY, ASK] Do you live with any family members?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

95. [IF NECESSARY, ASK] Do you live with people who are not family or are not related to you?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

Interviewer Questions

THE FOLLOWING QUESTIONS SHOULD BE ANSWERED AFTER THE INTERVIEW IS CONDUCTED.

96. WAS THE RESPONDENT ABLE TO GIVE VALID RESPONSES?

1☐ YES
2☐ NO

97. WAS ANY ONE ELSE PRESENT DURING THE INTERVIEW?

1☐ YES
2☐ NO → END SURVEY

98. WHO WAS PRESENT DURING THE INTERVIEW? (MARK ALL THAT APPLY.)

1☐ SOMEONE **NOT** PAID TO PROVIDE SUPPORT TO THE RESPONDENT
2☐ STAFF OR SOMEONE PAID TO PROVIDE SUPPORT TO THE RESPONDENT

99. DID SOMEONE HELP THE RESPONDENT COMPLETE THIS SURVEY?

1☐ YES
2☐ NO → END SURVEY

100. HOW DID THAT PERSON HELP? [MARK ALL THAT APPLY.]

1☐ ANSWERED ALL THE QUESTIONS FOR RESPONDENT
2☐ RESTATED THE QUESTIONS IN A DIFFERENT WAY OR REMINDED/ PROMPTED THE RESPONDENT
3☐ TRANSLATED THE QUESTIONS OR ANSWERS INTO THE RESPONDENT'S LANGUAGE
4☐ HELPED WITH THE USE OF ASSISTIVE OR COMMUNICATION EQUIPMENT SO THAT THE RESPONDENT COULD ANSWER THE QUESTIONS
5☐ OTHER, SPECIFY_____

101. WHO HELPED THE RESPONDENT? (MARK ALL THAT APPLY.)

1☐ SOMEONE **NOT** PAID TO PROVIDE SUPPORT TO THE RESPONDENT
2☐ STAFF OR SOMEONE PAID TO PROVIDE SUPPORT TO THE RESPONDENT

Supplemental Employment Module

IF THE SUPPLEMENTAL MODULE IS USED, IT SHOULD BE INSERTED PRIOR TO THE "ABOUT YOU" SECTION IN THE CORE SURVEY.

EM1. Do you work for pay at a job?

1☐ YES → GO TO EM9
2☐ NO
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM2. Do you want to work for pay at a job?

1☐ YES
2☐ NO → GO TO EM4
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM3. Sometimes people feel that something is holding them back from working when they want to. Is this true for you? If so, what is holding you back from working? (INTERVIEWER LISTENS AND MARKS ALL THAT APPLY)

1☐ BENEFITS → GO TO EM5
2☐ HEALTH CONCERNS → GO TO EM5
3☐ DON'T KNOW ABOUT JOB RESOURCES → GO TO EM5
4☐ ADVICE FROM OTHERS → GO TO EM5
5☐ TRAINING/EDUCATION NEED → GO TO EM5
6☐ LOOKING AND CAN'T FIND WORK → GO TO EM5
7☐ ISSUES WITH PREVIOUS EMPLOYMENT → GO TO EM5
8☐ TRANSPORTATION → GO TO EM5
9☐ CHILD CARE → GO TO EM5
10☐ OTHER (_____)→ GO TO EM5
11☐ NOTHING IS HOLDING ME BACK → GO TO EM5
-1☐ DON'T KNOW → GO TO EM5
-2☐ REFUSED → GO TO EM5
-3☐ UNCLEAR RESPONSE → GO TO EM5

EM4. Sometimes people would like to work for pay, but feel that something is holding them back. Is this true for you? If so, what is holding you back from wanting to work? (INTERVIEWER LISTENS AND MARKS ALL THAT APPLY)

1☐ BENEFITS → GO TO THE ABOUT YOU SECTION
2☐ HEALTH CONCERNS → GO TO THE ABOUT YOU SECTION
3☐ DON'T KNOW ABOUT JOB RESOURCES → GO TO THE ABOUT YOU SECTION
4☐ ADVICE FROM OTHERS → GO TO THE ABOUT YOU SECTION
5☐ TRAINING/EDUCATION NEED → GO TO THE ABOUT YOU SECTION
6☐ LOOKING AND CAN'T FIND WORK → GO TO THE ABOUT YOU SECTION
7☐ ISSUES WITH PREVIOUS EMPLOYMENT → GO TO THE GO TO THE ABOUT YOU SECTION
8☐ TRANSPORTATION → GO TO THE GO TO THE ABOUT YOU SECTION
9☐ CHILD CARE → GO TO THE ABOUT YOU SECTION
10☐ OTHER (_____)→ GO TO THE ABOUT YOU SECTION
11☐ NOTHING/DOESN'T WANT TO WORK → GO TO THE ABOUT YOU SECTION
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM5. Have you asked for help in getting a job for pay?

1☐ YES → GO TO EM7
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

EM6. Do you know you can get help to find a job for pay?

1☐ YES → GO TO THE ABOUT YOU SECTION
2☐ NO → GO TO THE ABOUT YOU SECTION
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM7. Help getting a job can include help finding a place to work or help getting the skills that you need to work. Is someone paid to help you get a job?

1☐ YES → GO TO EM8
2☐ NO → GO TO THE ABOUT YOU SECTION
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM8. Are you getting all the help you need to find a job?

1☐ YES → GO TO THE ABOUT YOU SECTION
2☐ NO → GO TO THE ABOUT YOU SECTION
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM9. Who helped you to find the job that you have now? [MARK ALL THAT APPLY]

1☐ EMPLOYMENT/VOCATIONAL STAFF/JOB COACH
2☐ CASE MANAGER
3☐ OTHER PAID PROVIDERS
4☐ OTHER CAREER SERVICES
5☐ FAMILY/FRIENDS
6☐ ADVERSTISEMENT
7☐ SELF-EMPLOYED → GO TO EM11
8☐ OTHER (_____)
9☐ NO ONE HELPED ME—I FOUND IT MYSELF → GO TO EM11
-1☐ DON'T KNOW → GO TO EM11
-2☐ REFUSED → GO TO EM11
-3☐ UNCLEAR RESPONSE → GO TO EM11

EM10. Did you help to choose the job you have now?

1☐ YES
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

EM11. Sometimes people need help from other people to work at their jobs. For example, they may need help getting to or getting around at work, help getting their work done, or help getting along with other workers. Is someone paid to help you with the job you have now?

1☐ YES
2☐ NO → GO TO THE ABOUT YOU SECTION
-1☐ DON'T KNOW→ GO TO THE ABOUT YOU SECTION
-2☐ REFUSED→ GO TO THE ABOUT YOU SECTION
-3☐ UNCLEAR RESPONSE→ GO TO THE ABOUT YOU SECTION

EM12. What do you call this person? A job coach, peer support provider, personal assistant, or something else?

_____

[USE THIS TERM WHEREVER IT SAYS {*job coach*} BELOW.]

EM13. Did you hire your {*job coach*} yourself?

1☐ YES → GO TO THE ABOUT YOU SECTION
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

EM14. Is your {*job coach*} with you all the time that you are working?

1☐ YES
2☐ NO
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

EM15. How often does your {*job coach*} give you all the help you need? Would you say . . .

1☐ Never,
2☐ Sometimes,
3☐ Usually, or
4☐ Always?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Does your {*job coach*} give you all the help you need? Would you say . . .

1☐ Mostly yes, or,
2☐ Mostly no?
-1☐ DON'T KNOW
-2☐ REFUSED
-3☐ UNCLEAR RESPONSE

EM16. How often is your {*job coach*} nice and polite to you? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Is your {*job coach*} nice and polite to you? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

EM17. How often does your {*job coach*} explain things in a way that is easy to understand? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Does your {*job coach*} explain things in a way that is easy to understand? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

EM18. How often does your {*job coach*} listen carefully to you? Would you say . . .

[1] ☐ Never,
[2] ☐ Sometimes,
[3] ☐ Usually, or
[4] ☐ Always?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: Does your {*job coach*} listen carefully to you? Would you say . . .

[1] ☐ Mostly yes, or,
[2] ☐ Mostly no?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

EM19. Does your {*job coach*} encourage you to do things for yourself if you can?

[1] ☐ YES
[2] ☐ NO
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

EM20. Using any number from 0 to 10, where 0 is the worst help from {*job coach*} possible and 10 is the best help from {*job coach*} possible, what number would you use to rate the help you get from {*job coach*}?

__0 TO 10

[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

ALTERNATE VERSION: How would you rate the help you get from your {*job coach*}? Would you say . . .

[1] ☐ Excellent,
[2] ☐ Very good,
[3] ☐ Good,
[4] ☐ Fair, or
[5] ☐ Poor?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

EM21. Would you recommend the {*job coach*} who helps you to your family and friends if they needed {*program-specific term for employment services*}? Would you say you recommend the {*job coach*} . . .

[1] ☐ Definitely no,
[2] ☐ Probably no,
[3] ☐ Probably yes, or
[4] ☐ Definitely yes?
[-1] ☐ DON'T KNOW
[-2] ☐ REFUSED
[-3] ☐ UNCLEAR RESPONSE

Home and Community Based Services Experience of Care Survey

Version: 1.0

Population: Adult

Language: Spanish

Response Scale: 4 point and 2 point alternative

Notes

Supplemental items: Survey users may add questions to this survey. The supplemental items are available at the end of this survey

Encuesta sobre las experiencias del usuario con los servicios que recibe en el hogar y la comunidad

Instructions for Vendor

- The interview is intended as an interviewer-administered survey, thus all text that appears in initial uppercase and lowercase letters should be read aloud. Text that appears in **bold, lowercase letters** should be emphasized.
- Text in {*italics and in braces*} will be provided by the HCBS program's administrative data. However, if the interviewee provides another term, that term should be used in place of the program-specific term wherever indicated. For example, some interviewees may refer to their case manager by another title, which should be used instead throughout the survey.
- For response options of "never, sometimes, usually, and always", if the respondent cannot use that scale, the alternate version of the survey should be used which uses the response options of "mostly yes and mostly no." These response options are reserved for individuals who find the "never, sometimes, usually, always" response scale cognitively challenging.
- For response options of 0 to 10, if the respondent cannot use that scale, the alternate version of the survey should be used which uses the response options of "Excellent," "very good," "good," "fair," or "poor." These response options are reserved for respondents who find the numeric scale cognitively challenging.
- All questions include a "REFUSED" response option. In this case, "refused" means the respondent did not provide any answer to the question.
- All questions include a "DON'T KNOW" response option. This is used when the respondent indicates that he or she does not know the answer and cannot provide a response to the question.
- All questions include an "UNCLEAR" response option. This should be used when a respondent answers, but the interviewer cannot clarify the meaning of the response even after minor probing or the response is completely unrelated to the question—for example, the response to "Do your homemakers listen carefully to what you say?" is "I like to sit by Mary."
- Some responses have skip patterns, which are expressed as "➔ GO TO Q #." The interviewer will be automatically skipped to the next correct item.
- Not all respondents have all services. Items Q4 through Q12 help to confirm which services a respondent has. The table after it presents the logic of which items should be used.
- **Use Singular/Plural as needed:** Modify items such that the interviewer can use the correct form (singular or plural) of the survey item.
- **Use Program-Specific Terms:** Where appropriate, add in the program-specific terms for staff (e.g., [*program-specific term for these types of staff*]) but allow the interviewer to modify the term based on the respondent's choice of the word. It will be necessary to obtain information for program-specific terms. State administrative data should include the following information:
  - Agency name(s)
  - Titles of staff who provide care
  - Names of staff who provide care
  - Activities that each staff member provides (this will help with identifying appropriate skip logic)
  - Hours of staff who come to the home

COGNITIVE SCREENING QUESTIONS

Es posible que a algunas personas se les pague para que le ayuden a alistarse por la mañana, a hacer los oficios de la casa, a ir a algún sitio o a recibir servicios de salud mental. Esta encuesta es sobre las personas a las que se les paga para que le ayuden con las actividades que hace normalmente o comúnmente en la casa y en la comunidad. También contiene preguntas sobre los servicios que recibe.

1. ¿Viene alguien a su casa para ayudarle?

   1 ☐ SÍ
   2 ☐ NO ➔ END SURVEY
   -1 ☐ NO SABE ➔ END SURVEY
   -2 ☐ SE NEGÓ A CONTESTAR ➔ END SURVEY
   -3 ☐ RESPUESTA POCO CLARA ➔ END SURVEY

2. ¿Cómo le ayudan?

   _____

   [EXAMPLES OF CORRECT RESPONSES INCLUDE]

   - ME AYUDA A ALISTARME TODOS LOS DIAS (HELPS ME GET READY EVERY DAY)
   - LIMPIA MI CASA (CLEANS MY HOME)
   - TRABAJA CONMIGO EN MI EMPLEO (WORKS WITH ME AT MY JOB)
   - ME AYUDA HACER COSAS (HELPS ME TO DO THINGS)
   - ME AYUDA CON TRANSPORTE (DRIVES ME AROUND)

   ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
   ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
   ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

3. ¿Cómo llama usted a esa(s) persona(s)?

   _____

   [EXAMPLES OF SUFFICIENT RESPONSES INCLUDE]

   - MI TRABAJADOR(A) (MY WORKER)
   - MI ASISTENTE (MY ASSISTANT)
   - POR SU(S) NOMBRE(S), MARIA, ANA, ETCC. (NAMES OF STAFF (JO, DAWN, ETC.))

   ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
   ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
   ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

CSQPASS.
(INT: IF ALL 3 QUESTIONS ANSWERED CORRECTLY, ENTER 1 TO CONTINUE.)

1 PASS - ALL 3 QUESTIONS WERE ANSWERED CORRECTLY ➔ GO TO Q4
2 FAIL - AT LEAST 1 QUESTION WAS NOT ANSWERED CORRECTLY ➔ GO TO SURVEND

SURVEND.
Gracias por su tiempo. Esas son todas la preguntas que tenemos.
Tenga un buen día/tarde. (INT: ENTER 1 TO EXIT SURVEY)

PREGUNTAS DE IDENTIFICACIÓN

Ahora me gustaría hacerle más preguntas sobre el tipo de personas que vienen a su casa para ayudarle.

4. ¿Recibe usted {*program specific term for personal assistance*} en casa?

   1 ☐ SÍ
   2 ☐ NO ➔ Go to Q6
   -1 ☐ NO SABE ➔ Go to Q6
   -2 ☐ SE NEGÓ A CONTESTAR ➔ Go to Q6
   -3 ☐ RESPUESTA POCO CLARA ➔ Go to Q6

5. ¿Cómo llama usted a la(s) persona(s) que le da(n) {*program specific term for personal assistance*}? Por ejemplo, ¿les llama {*program specific term for personal assistance}*, personal, auxiliares de cuidados personales (*PCAs* por su sigla en inglés), trabajadores o alguna otra cosa?

   _____

   [ADD RESPONSE WHEREVER IT SAYS "*personal assistance/behavioral health staff", "el personal de salud mental / los auxiliares de cuidados personales*"]

6. ¿Recibe usted {*program specific term for behavioral health specialist services*} en casa?

   1 ☐ SÍ
   2 ☐ NO ➔ Go to Q8
   -1 ☐ NO SABE ➔ Go to Q8
   -2 ☐ SE NEGÓ A CONTESTAR ➔ Go to Q8
   -3 ☐ RESPUESTA POCO CLARA ➔ Go to Q8

7. ¿Cómo llama usted a la(s) persona(s) que le da(n) {*program specific term for behavioral health specialist services*}? Por ejemplo, ¿les llama {*program specific term for behavioral health specialists*}, consejeros, apoyo de personas en la misma situación (*peer support* en inglés), asistentes de recuperación o alguna otra cosa?

   _____

   [ADD RESPONSE WHEREVER IT SAYS "*personal assistance/behavioral health staff";* IF Q4 IS ALSO= YES, LIST BOTH TITLES]

8. ¿Recibe usted {*program specific term for homemaker services*} en casa?

   1 ☐ SÍ
   2 ☐ NO ➔ GO to Q11
   -1 ☐ NO SABE ➔ Q11
   -2 ☐ SE NEGÓ A CONTESTAR ➔ GO to Q11
   -3 ☐ RESPUESTA POCO CLARA ➔ GO to Q11

9. ¿Cómo llama usted a la(s) persona(s) que le da(n) {*program specific term for homemaker services*}? Por ejemplo, ¿les llama {*program specific term for homemaker*}, ayudantes de oficios domésticos, ayudantes para tareas de la casa o alguna otra cosa?

_____

[ADD RESPONSE WHEREVER IT SAYS "*homemaker"*]

10. [IF (Q4 *OR* Q6) *AND* Q8= YES, ASK] Las personas que le ayudan con las actividades que hace normalmente o comúnmente ¿también le ayudan a limpiar la casa?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

11. ¿Recibe usted ayuda de {*program specific term for case manager services*} para asegurarse de que usted reciba todos los servicios que necesita?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

12. ¿Cómo llama usted a la persona que le da {*program specific term for case manager services*}? Por ejemplo, ¿llama a esa persona {*program specific term for case manager*}, encargado de caso, encargado de cuidados, coordinador de servicios, coordinador de servicios de apoyo, trabajador social o alguna otra cosa?

_____

[ADD RESPONSE WHEREVER IT SAYS "*case manager"*]

BELOW ARE INSTRUCTIONS TO WHICH QUESTIONS TO ASK FOR EACH RESPONSE ABOVE

ITEM AND RESPONSE	**ACTION**
IF Q4 OR Q6= YES, AND Q8 = NO, DON'T KNOW, REFUSE, UNCLEAR	ASK Q13-Q36, AND Q48 ONWARD
IF Q4 AND Q6 = NO	SKIP Q13-36, 57 AND 79
IF Q8 = YES	ASK Q37-Q47, AND Q48 ONWARD
IF Q10 = YES	ASK Q13-Q36, Q39, Q40, AND Q48 ONWARD
IF Q11 = ANY RESPONSE	ASK Q48 – Q55, AND Q56 ONWARD

Obtención de los servicios necesarios de parte de los auxiliares de cuidados personales y del personal de salud mental

13. Primero me gustaría hablar sobre {*el personal de salud mental / los auxiliares de cuidados personales*} la(s) persona(s) a la(s) que se le(s) paga para que le ayude(n) en sus actividades diarias, como vestirse, ir al baño, bañarse o ducharse, o ir a algún sitio. ¿Con qué frecuencia el/los {*el personal de salud mental / los auxiliares de cuidados personales*} llega(n) a trabajar a tiempo? ¿Diría que…?

- [1] ☐ Nunca,
- [2] ☐ A veces,
- [3] ☐ Casi siempre, o
- [4] ☐ Siempre?
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: Primero me gustaría hablar sobre el/los{*el personal de salud mental / los auxiliares de cuidados personales*}, a quien(es) se le(s) paga para que le ayude(n) en sus actividades diarias, como vestirse, ir al baño, bañarse o ducharse, o ir a algún sitio. ¿Llega(n) el/los {*el personal de salud mental / los auxiliares de cuidados personales*} a trabajar a tiempo? ¿Diría que…?

- [1] ☐ En general, sí, o
- [2] ☐ En general, no?
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

14. ¿Con qué frecuencia {*el personal de salud mental / los auxiliares de cuidados personales*} trabaja(n) todo el tiempo que se supone que debe(n) trabajar? ¿Diría que…?

- [1] ☐ Nunca,
- [2] ☐ A veces,
- [3] ☐ Casi siempre, o
- [4] ☐ Siempre?
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Trabaja(n) el/los {*el personal de salud mental / los auxiliares de cuidados personales*} todo el tiempo que se supone que debe(n) trabajar? ¿Diría que…?

- [1] ☐ En general, sí, o
- [2] ☐ En general, no?
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

15. A veces el personal no puede llegar al trabajo en un día en que tenga programado hacerlo. Cuando el personal no puede llegar al trabajo en un día en que tenga programado hacerlo, ¿le avisa alguien si {*el personal de salud mental / los auxiliares de cuidados personales*} no puede llegar ese día?

- [1] ☐ SÍ
- [2] ☐ NO

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

16. ¿Necesita ayuda de {*el personal de salud mental / los auxiliares de cuidados personales*} para vestirse, ducharse o bañarse?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q20
-1 ☐ NO SABE ➔ GO TO Q20
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q20
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q20

17. ¿**Siempre** se viste, se ducha o se baña cuando lo necesita?

1 ☐ SÍ ➔ GO TO Q19
2 ☐ NO
-1 ☐ NO SABE ➔ GO TO Q19
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q19
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q19

18. ¿Esto pasa porque no hay {*auxiliares de cuidados personales / personal de salud mental*} que le ayude(n)?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

19. ¿Con qué frecuencia {*el personal de salud mental / los auxiliares de cuidados personales*} se asegura(n) de que usted tenga suficiente privacidad cuando se viste, se ducha o se baña? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Se asegura(n) el/los {*el personal de salud mental / los auxiliares de cuidados personales*} de que usted tenga suficiente privacidad cuando se viste, se ducha o se baña? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

20. ¿Necesita que {*el personal de salud mental / los auxiliares de cuidados personales*} le ayude(n) con las comidas, por ejemplo, para preparar o cocinar las comidas o para ayudarle a comer?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q23

-1 ☐ NO SABE ➔ GO TO Q23
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q23
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q23

21. ¿**Siempre** puede conseguir algo para comer cuando tiene hambre?

1 ☐ SÍ ➔ GO TO Q23
2 ☐ NO
-1 ☐ NO SABE ➔ GO TO Q23
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q23
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q23

22. ¿Esto pasa porque no hay {*auxiliares de cuidados personales / personal de salud mental*} que le ayude(n)?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

23. A veces las personas necesitan ayuda para tomarse sus medicinas, por ejemplo, necesitan ayuda para acordarse de tomárselas, para servirlas o para alistar las pastillas. ¿Necesita que {*el personal de salud mental / los auxiliares de cuidados personales*} le ayude(n) a tomarse sus medicinas?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q26
-1 ☐ NO SABE ➔ GO TO Q26
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q26
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q26

24. ¿**Siempre** se toma su medicina cuando debe tomársela?

1 ☐ SÍ ➔ GO TO Q26
2 ☐ NO
-1 ☐ NO SABE ➔ GO TO Q26
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q26
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q26

25. ¿Esto pasa porque no hay {*auxiliares de cuidados personales / personal de salud mental*} que le ayude(n)?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

26. La ayuda para ir al baño incluye ayudarle a alguien a sentarse y levantarse del inodoro o ayudarle a cambiarse de ropa interior o de toallas desechables. ¿Necesita que {*el personal de salud mental / los auxiliares de cuidados personales*} le ayude(n) a ir al baño?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q28
-1 ☐ NO SABE ➔ GO TO Q28
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q28
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q28

27. ¿Recibe usted toda la ayuda que necesita de {*el personal de salud mental / los auxiliares de cuidados personales*} para ir al baño cuando lo necesita?

[1] ☐ SÍ
[2] ☐ NO
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Qué tan bien se comunica(n) con usted los auxiliares de cuidados personales o el personal de salud mental y qué tan bien lo(a) tratan

Las siguientes preguntas se refieren a cómo lo(a) trata(n) {*el personal de salud mental / los auxiliares de cuidados personales*}.

28. ¿Con qué frecuencia {*el personal de salud mental / los auxiliares de cuidados personales*} es/son amable(s) y educado(s) con usted? ¿Diría que…?

[1] ☐ Nunca,
[2] ☐ A veces,
[3] ☐ Casi siempre, o
[4] ☐ Siempre?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*El personal de salud mental / los auxiliares de cuidados personales*} es/son amable(s) y educado(s) con usted? ¿Diría que…?

[1] ☐ En general, sí, o
[2] ☐ En general, no?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

29. ¿Con qué frecuencia es difícil entender las explicaciones que le da(n) {*el personal de salud mental / los auxiliares de cuidados personales*} porque tiene(n) acento o por la forma en que ellos o ellas hablan español? ¿Diría que…?

[1] ☐ Nunca,
[2] ☐ A veces,
[3] ☐ Casi siempre, o
[4] ☐ Siempre?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Es difícil entender las explicaciones que le da(n) {*el personal de salud mental / los auxiliares de cuidados personales*} porque estos tiene(n) acento o por la forma en que hablan español? ¿Diría que…?

[1] ☐ En general, sí, o
[2] ☐ En general, no?
[-1] ☐ NO SABE

-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

30. ¿Con qué frecuencia {l*os auxiliares de cuidados personales / el personal de salud mental*} lo(a) trata(n) como usted quiere? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*El personal de salud mental / los auxiliares de cuidados personales*} lo(a) trata(n) como usted quiere? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

31. ¿Con qué frecuencia {*el personal de salud mental / los auxiliares de cuidados personales*} le explica(n) las cosas de una manera fácil de entender? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*El personal de salud mental / los auxiliares de cuidados personales*} le explica(n) las cosas de una manera fácil de entender? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

32. ¿Con qué frecuencia {*el personal de salud mental / los auxiliares de cuidados personales*} lo(a) escuchan con atención? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*El personal de salud mental / los auxiliares de cuidados personales*} lo(a) escucha(n) con atención? ¿Diría que…?

[1] ☐ En general, sí, o
[2] ☐ En general, no?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

33. ¿Cree usted que {*el personal de salud mental / los auxiliares de cuidados personales*} sabe(n) el tipo de ayuda que **usted** necesita con las actividades diarias, como alistarse por la mañana, hacer mercado o ir a alguna parte de su comunidad?

[1] ☐ SÍ
[2] ☐ NO
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

34. ¿{*El personal de salud mental / los auxiliares de cuidados personales*} lo(a) animan a hacer cosas sin ayuda si usted puede hacerlas?

[1] ☐ SÍ
[2] ☐ NO
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

35. Usando un número del 0 al 10, el 0 siendo la peor ayuda que recibe de {*el personal de salud mental / los auxiliares de cuidados personales*} posible y el 10 es la mejor ayuda que recibe de {*el personal de salud mental / los auxiliares de cuidados personales*} posible, ¿qué número usaría para calificar la ayuda que recibe de {*el personal de salud mental / los auxiliares de cuidados personales*}?

___ 0 a 10

[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Cómo calificaría la ayuda que recibe de {*el personal de salud mental / los auxiliares de cuidados personales*}? ¿Diría que es…?

[1] ☐ Excelente,
[2] ☐ Muy buena,
[3] ☐ Buena,
[4] ☐ Regular, o
[5] ☐ Mala?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

36. ¿Les recomendaría a sus familiares y amigos {*el personal de salud mental / los auxiliares de cuidados personales*} que le ayuda(n) si ellos necesitaran ayuda para realizar las actividades diarias? ¿Diría que recomendaría {*el personal de salud mental / los auxiliares de cuidados personales*}?

1 ☐ Definitivamente no,
2 ☐ Probablemente no,
3 ☐ Probablemente sí, o
4 ☐ Definitivamente sí?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Obtención de los servicios necesarios de los ayudantes de oficios domésticos

Las siguientes preguntas son acerca de los {*ayudantes de oficios domésticos*}, el personal a quien se le paga para que haga tareas de la casa, como limpiar, hacer mercado o lavar la ropa.

37. ¿Con qué frecuencia {*los ayudantes de oficios domésticos*} llegan a tiempo al trabajo? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Llegan los {*los ayudantes de oficios domésticos*} a tiempo al trabajo? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

38. ¿Con qué frecuencia {l*os ayudantes de oficios domésticos*} trabajan todo el tiempo que se supone que deben trabajar? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Trabajan {l*os ayudantes de oficios domésticos*} todo el tiempo que se supone que deben trabajar? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR

-3 ☐ RESPUESTA POCO CLARA

39. Las tareas de la casa, como limpiar y lavar la ropa ¿se hacen **siempre** cuando usted necesita que se hagan? [ASK IF HOMEMAKER IS THE SAME AS PCA STAFF]

1 ☐ SÍ ➔ GO TO Q41
2 ☐ NO
-1 ☐ NO SABE ➔ GO TO Q41
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q41
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q41

40. ¿Es porque no hay {*ayudantes de oficios domésticos*} que le ayuden? [ASK IF HOMEMAKER IS THE SAME AS PCA STAFF]

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Qué tan bien se comunican con usted los ayudantes de oficios domésticos y qué tan bien lo(a) tratan

Las siguientes preguntas se refieren a la forma en que lo(a) tratan *{los ayudantes de oficios domésticos}*.

41. ¿Con qué frecuencia {*los ayudantes de oficios domésticos*} son amables y educados con usted? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa:¿{*Los ayudantes de oficios domésticos*} son amables y educados con usted? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

42. ¿Con qué frecuencia es difícil entender las explicaciones que le dan {*los ayudantes de oficios domésticos*} porque tienen acento o por la forma en que ellos o ellas hablan español? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE

-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Es difícil entender las explicaciones que le dan {*los ayudantes de oficios domésticos*} porque estos tienen acento o por la forma en que {*los ayudantes de oficios domésticos*} hablan español? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

43. ¿Con qué frecuencia {*los ayudantes de oficios domésticos*} lo(a) tratan como usted quiere? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*Los ayudantes de oficios domésticos*} lo(a) tratan como usted quiere? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

44. ¿Con qué frecuencia {*Los ayudantes de oficios domésticos*} le escuchan con atención? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿{*Los ayudantes de oficios domésticos*} le escuchan con atención? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

45. ¿Cree usted que {*los ayudantes de oficios domésticos*} saben el tipo de ayuda que usted necesita?

1 ☐ SÍ
2 ☐ NO

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

46. ¿Usando un número del 0 al 10, el 0 siendo la peor ayuda que recibe de {*los ayudantes de oficios domésticos*} posible y el 10 es la mejor ayuda que recibe de {*los ayudantes de oficios domésticos*} posible, ¿qué número usaría para calificar la ayuda que recibe de {*los ayudantes de oficios domésticos*}?

___ 0 a 10

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Cómo calificaría la ayuda que recibe de {*los ayudantes de oficios domésticos*}? ¿Diría que es…?

1 ☐ Excelente,
2 ☐ Muy buena,
3 ☐ Buena,
4 ☐ Regular, o
5 ☐ Mala?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

47. ¿Les recomendaría a sus familiares y amigos {*los ayudantes de oficios domésticos*} que le ayudan si ellos necesitaran {*término específico del encuestado para "servicios de ayuda con los oficios domésticos"*}? ¿Diría que recomendaría {los ayudantes de oficios domésticos}?

1 ☐ Definitivamente no,
2 ☐ Probablemente no,
3 ☐ Probablemente sí, o
4 ☐ Definitivamente sí?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Su encargado de caso

Ahora me gustaría hablarle de su *{encargado de caso}*, la persona que se asegura de que usted reciba los servicios que necesita.

48. ¿Sabe quién es su {*encargado de caso*}?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q56
-1 ☐ NO SABE ➔ GO TO Q56
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q56
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q56

49. ¿Puede comunicarse con este {*encargado de caso*} cuando necesita hacerlo?

1 ☐ SÍ
2 ☐ NO

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

50. Algunas personas necesitan conseguir equipo, como sillas de ruedas o andadores, que les sirvan de ayuda y otras personas necesitan que el equipo que tienen sea remplazado o reparado. ¿Le ha pedido ayuda a este {*encargado de caso*} para conseguir o reparar un equipo?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q52
3 ☐ NO NECESITA➔ GO TO Q52
-1 ☐ NO SABE ➔ GO TO Q52
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q52
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q52

51. ¿Este {*encargado de caso*} colaboró con usted cuando le pidió ayuda para conseguir o reparar un equipo?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

52. ¿Le ha pedido ayuda a este {*encargado de caso*} para hacer cambios en los servicios que recibe, como más ayuda de {*el personal de salud mental/los auxiliares de cuidados personales y/o los ayudantes de oficios domésticos*}, o para ir a lugares o buscar trabajo?

1 ☐ SÍ
2 ☐ NO ➔ GO TO 54
3 ☐ NO NECESITA➔ GO TO Q54
-1 ☐ NO SABE ➔ GO TO 54
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO 54
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO 54

53. ¿Este {*encargado de caso*} colaboró con usted cuando le pidió ayuda para hacer otros cambios en los servicios que recibe?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

54. ¿Usando un número del 0 al 10, el 0 siendo la peor ayuda que recibe del {*encargado de caso*} posible y el 10 es la mejor ayuda que recibe del {*encargado de caso*} posible, ¿qué número usaría para calificar la ayuda que recibe del {*encargado de caso*}?

___ 0 a 10

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Cómo calificaría la ayuda que recibe del {*encargado de caso*}? ¿Diría que es…?

1 ☐ Excelente,
2 ☐ Muy buena,
3 ☐ Buena,
4 ☐ Regular, o
5 ☐ Mala?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

55. ¿Les recomendaría a sus familiares y amigos el {*encargado de caso*} que le ayuda a usted si ellos necesitaran {*término específico del encuestado para "servicios que presta un encargado de caso"*}? ¿Diría que les recomendaría el {*encargado de caso*}?

1 ☐ Definitivamente no,
2 ☐ Probablemente no,
3 ☐ Probablemente sí, o
4 ☐ Definitivamente sí?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

La elección de sus servicios

56. ¿Qué se incluye en su [*término específico de cada programa que se refiere a un "plan de servicios"*]?

- [1] ☐ **Ninguna** de las cosas que son importantes para usted
- [2] ☐ **Algunas** de las cosas que son importantes para usted
- [3] ☐ **La mayoría** de las cosas que son importantes para usted
- [4] ☐ **Todas** las cosas que son importantes para usted
- [-1] ☐ NO SABE ➔ GO TO Q58
- [-2] ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q58
- [-3] ☐ RESPUESTA POCO CLARA ➔ GO TO Q58

57. ¿Cree que {*los auxiliares de cuidados personales / el personal de salud mental*} sabe(n) qué se incluye en su [*término específico de cada programa que se refiere a un "plan de servicios"*], incluso las cosas que son importantes para usted?

- [1] ☐ SÍ
- [2] ☐ NO
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

58. ¿Con quién hablaría si quisiera cambiar su [*término específico de cada programa que se refiere a un "plan de servicios*"]? ¿Hablaría con alguien más? [INTERVIEWER MARKS ALL THAT APPLY]

- [1] ☐ ENCARGADO DE CASO
- [2] ☐ OTROS MIEMBROS DEL PERSONAL
- [3] ☐ FAMILIARES/ AMIGOS
- [4] ☐ ALGUIEN MÁS, ESPECIFIQUE _____
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

Transporte

El tema de las siguientes preguntas es cómo va usted a sitios de su comunidad.

59. Entre las citas médicas se incluye ir a ver al doctor, al dentista, al terapeuta o a otra persona que se encargue del cuidado de su salud. ¿Con qué frecuencia tiene forma de llegar a sus citas médicas? ¿Diría que...?

- [1] ☐ Nunca,
- [2] ☐ A veces,
- [3] ☐ Casi siempre, o
- [4] ☐ Siempre?
- [-1] ☐ NO SABE
- [-2] ☐ SE NEGÓ A CONTESTAR
- [-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: Entre las citas médicas se incluye ir a ver al doctor, al dentista, al terapeuta o a otra persona que se encargue del cuidado de su salud. ¿Tiene forma de llegar a sus citas médicas? ¿Diría que...?

- [1] ☐ En general, sí, o
- [2] ☐ En general, no?

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

60. ¿Usa una camioneta van o algún otro servicio de transporte? No incluya una camioneta van que le pertenezca a usted.

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q63
-1 ☐ NO SABE ➔ GO TO Q63
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q63
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q63

61. ¿Puede subirse y bajarse de este vehículo fácilmente?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

62. ¿Con qué frecuencia llega este vehículo a tiempo a recogerlo(a)? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Llega este vehículo a tiempo a recogerlo(a)? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Seguridad personal

Las siguientes preguntas se refieren a su seguridad personal.

63. ¿Con quién se comunicaría en caso de emergencia? [INTERVIEWER MARKS ALL THAT APPLY]

1 ☐ PARIENTE O AMIGO
2 ☐ ENCARGADO DE CASO
3 ☐ AGENCIA
4 ☐ SERVICIO DE EMERGENCIA PAGADOS (EJEMPLO LIFELINE)
5 ☐ 911/ PERSONAL DE PRIMEROS AUXILIOS (POLICIA, ETC)
6 ☐ ALGUIEN MÁS, ESPECIFIQUE _____
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

64. ¿Hay una persona con quien pueda hablar si alguien lo(a) lastima o le hace algo que a usted no le gusta?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Las siguientes preguntas se refieren a si alguna persona a quien se le paga para ayudarle en este momento lo(a) está tratando mal. Esto incluye a *{personal assistance/behavioral health staff, homemakers, or your case manager}.* Les estamos haciendo a todos las siguientes preguntas, no solo a usted. [ADD STATE-SPECIFIC LANGUAGE HERE REGARDING MANDATED REPORTING, IF APPROPRIATE: Quiero recordarle que, aunque sus respuestas son confidenciales, tengo la responsabilidad legal de informarle al estado de *{STATE}* si oigo algo que me haga pensar que alguien lo(a) está lastimando o que usted está en peligro.]

65. ¿**Alguno** de {*los auxiliar(es) de cuidados personales, el personal de salud mental, los ayudantes de oficios domésticos o los encargados de caso*} que tiene **ahora** toma su dinero o sus cosas sin preguntarle primero?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q68
-1 ☐ NO SABE ➔ GO TO Q68
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO 68
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q68

66. ¿Alguien está colaborando con usted para solucionar este problema?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q68
-1 ☐ NO SABE ➔ GO TO Q68
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q68
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q68

67. ¿Quién está colaborando con usted para solucionar este problema? ¿Alguna otra persona? [INTERVIEWER MARKS ALL THAT APPLY]

1 ☐ PARIENTE O AMIGO
2 ☐ ENCARGADO DE CASO
3 ☐ AGENCIA
4 ☐ ALGUIEN MÁS, ESPECIFIQUE _____
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

68. ¿Algún {*empleado*} de los que tiene ahora le grita, lo(a) insulta o le dice malas palabras?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q71
-1 ☐ NO SABE ➔ GO TO Q71
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q71
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q71

69. ¿Alguien está colaborando con usted para solucionar este problema?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q71

-1 ☐ NO SABE ➔ GO TO Q71
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q71
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q71

70. ¿Quién está colaborando con usted para solucionar este problema? ¿Alguna otra persona? [INTERVIEWER MARKS ALL THAT APPLY]

1 ☐ PARIENTE O AMIGO
2 ☐ ENCARGADO DE CASO
3 ☐ AGENCIA
4 ☐ ALGUIEN MÁS, ESPECIFIQUE _____
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

71. ¿Algún {*empleado*} de los que tiene ahora le pega o lo(a) lastima?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q74
-1 ☐ NO SABE ➔ GO TO Q74
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q74
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q74

72. ¿Alguien está colaborando con usted para solucionar este problema?

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q74
-1 ☐ NO SABE ➔ GO TO Q74
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q74
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q74

73. ¿Quién está colaborando con usted para solucionar este problema? ¿Alguna otra persona? [INTERVIEWER MARKS ALL THAT APPLY]

1 ☐ PARIENTE O AMIGO
2 ☐ ENCARGADO DE CASO
3 ☐ AGENCIA
4 ☐ ALGUIEN MÁS, ESPECIFIQUE _____
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Comunidad y empoderamiento

Ahora me gustaría preguntarle sobre las cosas que hace en su comunidad.

74. ¿Tiene **familiares** que vivan cerca? No incluya a los miembros de la familia con los que vive.

1 ☐ SÍ
2 ☐ NO ➔ GO TO Q76
-1 ☐ NO SABE ➔ GO TO Q76
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q76
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO Q76

75. Cuando usted lo desea, ¿con qué frecuencia puede reunirse con estos familiares que viven cerca? ¿Diría que…?

[1] ☐ Nunca,
[2] ☐ A veces,
[3] ☐ Casi siempre, o
[4] ☐ Siempre?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: Cuando usted lo desea, ¿puede reunirse con estos familiares que viven cerca? ¿Diría que…?

[1] ☐ En general, sí, o
[2] ☐ En general, no?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

76. ¿Tiene **amigos** que vivan cerca?

[1] ☐ SÍ
[2] ☐ NO ➔ GO TO Q78
[-1] ☐ NO SABE ➔ GO TO Q78
[-2] ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q78
[-3] ☐ RESPUESTA POCO CLARA ➔ GO TO Q78

77. Cuando usted lo desea, ¿con qué frecuencia puede reunirse con estos amigos que viven cerca? ¿Diría que…?

[1] ☐ Nunca,
[2] ☐ A veces,
[3] ☐ Casi siempre, o
[4] ☐ Siempre?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

Versión Alternativa: Cuando usted lo desea, ¿puede reunirse con estos amigos que viven cerca? ¿Diría que…?

[1] ☐ En general, sí, o
[2] ☐ En general, no?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

78. Cuando usted lo desea, ¿con qué frecuencia puede hacer lo que le gusta en la comunidad?

[1] ☐ Nunca,
[2] ☐ A veces,
[3] ☐ Casi siempre, o
[4] ☐ Siempre?
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR

-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: Cuando usted lo desea, ¿puede hacer lo que le gusta en la comunidad?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

79. ¿Necesita más ayuda de la que recibe ahora de {*el personal de salud mental / los auxiliares de cuidados personales}* para hacer cosas en su comunidad?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

80. ¿Participa en decidir **qué** hace cada día?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

81. ¿Participa en decidir el **horario** de sus actividades de cada día? Por ejemplo, cuándo se levanta, cuándo come o cuándo se acuesta.

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Sobre usted

Ahora tengo unas cuantas preguntas sobre usted.

82. En general, ¿cómo calificaría toda su salud? ¿Diría que…?

1 ☐ Excelente,
2 ☐ Muy buena,
3 ☐ Buena,
4 ☐ Regular, o
5 ☐ Mala?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

83. En general, ¿cómo calificaría toda su salud mental o emocional? ¿Diría que…?

1 ☐ Excelente,
2 ☐ Muy buena,

3 ☐ Buena,
4 ☐ Regular, o
5 ☐ Mala?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

84. ¿Qué edad tiene?

1 ☐ ENTRE 18 Y 24 AÑOS ➔ GO TO Q85
2 ☐ ENTRE 25 Y 34 AÑOS ➔ GO TO Q85
3 ☐ ENTRE 35 Y 44 AÑOS ➔ GO TO Q85
4 ☐ ENTRE 45 Y 54 AÑOS ➔ GO TO Q85
5 ☐ ENTRE 55 Y 64 AÑOS ➔ GO TO Q85
6 ☐ ENTRE 65 Y 74 AÑOS ➔ GO TO Q85
7 ☐ 75 AÑOS O MÁS ➔ GO TO Q85
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q85
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿En qué año nació?

_____ (AÑO)

-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

85. [IF NECESSARY, ASK, AND VERIFY IF OVER THE PHONE] ¿Es usted hombre o mujer?

1 ☐ Hombre
2 ☐ Mujer
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

86. ¿Es de origen o ascendencia hispana o latina o española?

1 ☐ Sí, hispano(A), latino(A) o ESPAÑOL(A)
2 ☐ No, ni hispano(A) ni latino(A) ni ESPAÑOL(A) ➔ GO TO Q88
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

87. ¿Qué grupo lo describe mejor? [READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.]

1 ☐ Mexicano, mexicano americano, chicano
2 ☐ Puertorriqueño
3 ☐ Cubano
4 ☐ De otro origen hispano, latino o español
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

88. ¿A qué raza pertenece? Puede escoger una o más de las siguientes. ¿Diría que es…?

[1] ☐ Blanco(a) ➔ GO TO Q91
[2] ☐ Negro(a) o afroamericano(a) ➔ GO TO Q91
[3] ☐ Asiático(a) ➔ GO TO Q89
[4] ☐ Nativo(a) de Hawái o de otras islas del Pacifico ➔ GO TO Q90
[5] ☐ Indígena americano(a) o nativo(a) de Alaska ➔ GO TO 91
[6] ☐ OTRO ➔ GO TO Q91
[-1] ☐ NO SABE ➔ GO TO Q91
[-2] ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q91
[-3] ☐ RESPUESTA POCO CLARA ➔ GO TO Q91

89. ¿Qué grupo lo describe mejor? [READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.]

[1] ☐ Indio asiático ➔ GO TO Q91
[2] ☐ Chino ➔ GO TO Q91
[3] ☐ Filipino ➔ GO TO Q91
[4] ☐ Japonés ➔ GO TO Q91
[5] ☐ Coreano ➔ GO TO Q91
[6] ☐ Vietnamita ➔ GO TO Q91
[7] ☐ Otra asiático ➔ GO TO Q91
[-1] ☐ NO SABE ➔ GO TO Q91
[-2] ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q91
[-3] ☐ RESPUESTA POCO CLARA ➔ GO TO Q91

90. ¿Qué grupo lo describe mejor? READ ALL ANSWER CHOICES. CODE ALL THAT APPLY.

[1] ☐ Nativo de Hawái ➔ GO TO Q91
[2] ☐ Guameño o chamorro ➔ GO TO Q91
[3] ☐ Samoano ➔ GO TO Q91
[4] ☐ Nativa de otras islas del Pacífico ➔ GO TO Q91
[-1] ☐ NO SABE ➔ GO TO Q91
[-2] ☐ SE NEGÓ A CONTESTAR ➔ GO TO Q91
[-3] ☐ RESPUESTA POCO CLARA ➔ GO TO Q91

91. ¿Habla algún otro idioma aparte de español en casa? READ CHOICES ONLY IF NEEDED…

[1] ☐ SÍ
[2] ☐ No ➔ [GO TO Q93]
[-1] ☐ NO SABE ➔ [GO TO Q93]
[-2] ☐ SE NEGÓ A CONTESTAR ➔ [GO TO Q93]
[-3] ☐ RESPUESTA POCO CLARA ➔ [GO TO Q93]

92. ¿Qué idioma habla usted en casa?

[1] ☐ Inglés
[2] ☐ Español
[3] ☐ Ambos: inglés y español
[4] ☐ Español y otro idioma
[5] ☐ Otro idioma ➔ ¿Cuál? _____
[-1] ☐ NO SABE
[-2] ☐ SE NEGÓ A CONTESTAR
[-3] ☐ RESPUESTA POCO CLARA

93. **[**IF NECESSARY, ASK**]** ¿Incluyendo a usted, cuantos adultos viven en su casa?

1 ☐ 1 [SOLO EL RESPONDENTE] ➔ END SURVEY
2 ☐ ENTRE 2 A 3
3 ☐ 4 O MÁS
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

94. **[**IF NECESSARY, ASK**]** ¿Vive con familiares?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

95. [IF NECESSARY, ASK] ¿Vive con personas que no son de su familia ni tienen ningún parentesco con usted?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Interviewer Questions

LAS SIGUIENTES PREGUNTAS DBERAN CONTESTARSE DESPUES DE LA ENTREVISTA.

96. ¿EL ENTREVISTADO PUDO DAR RESPUESTAS VÁLIDAS?

1 ☐ SÍ
2 ☐ NO

97. ¿ESTUVO ALGUNA OTRA PERSONA PRESENTE DURANTE LA ENTREVISTA?

1 ☐ SÍ
2 ☐ NO ➔ END SURVEY

98. ¿QUIÉN ESTUVO PRESENTE DURANTE LA ENTREVISTA? (MARQUE TODAS LAS OPCIONES QUE CORRESPONDAN)

1 ☐ ALGUIEN A QUIEN **NO** SE LE PAGA PARA QUE LE PROPORCIONE APOYO AL ENTREVISTADO
2 ☐ UN MIEMBRO DEL PERSONAL O ALGUIEN A QUIEN SE LE PAGA PARA QUE LE PROPORCIONE APOYO AL ENTREVISTADO

99. ¿ALGUIEN LE AYUDÓ AL ENTREVISTADO A RESPONDER ESTA ENCUESTA?

1 ☐ SÍ
2 ☐ NO ➔ END SURVEY

100. ¿CÓMO LE AYUDÓ ESA PERSONA? MARQUE TODAS LAS OPCIONES QUE CORRESPONDAN.

1 ☐ RESPONDIÓ A TODAS LAS PREGUNTAS POR EL ENTREVISTADO
2 ☐ FORMULÓ LAS PREGUNTAS DE DIFERENTE MANERA O LE RECORDÓ / LE DIO PISTAS AL ENTREVISTADO
3 ☐ TRADUJO LAS PREGUNTAS O RESPUESTAS AL IDIOMA DEL ENTREVISTADO
4 ☐ AYUDÓ MEDIANTE EL USO DE UN EQUIPO DE ASISTENCIA O DE COMUNICACIONES PARA QUE EL ENTREVISTADO PUDIERA RESPONDER A LAS PREGUNTAS
5 ☐ OTRA (ESPECIFIQUE) _____

101. ¿QUIÉN LE AYUDÓ AL ENTREVISTADO? (MARQUE TODAS LAS OPCIONES QUE CORRESPONDAN)

1 ☐ ALGUIEN A QUIEN **NO** SE LE PAGA PARA QUE LE PROPORCIONE APOYO AL ENTREVISTADO
2 ☐ UN MIEMBRO DEL PERSONAL O ALGUIEN A QUIEN SE LE PAGA PARA QUE LE PROPORCIONE APOYO AL ENTREVISTADO

MÓDULO COMPLEMENTARIO SOBRE EMPLEO

SI SE USA EL MÓDULO COMPLEMENTARIO, INSERTELO ANTES DE LA SECCION "SOBRE USTED" EN LA ENCUESTA.

EM1. ¿Tiene un trabajo por el cual le pagan?

1 ☐ SÍ ➔ GO TO QEM9
2 ☐ NO
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM2. ¿Quiere tener un trabajo por el cual le paguen?

1 ☐ SÍ
2 ☐ NO ➔ GO TO EM4
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM3. A veces, uno cree que hay algo que le impide trabajar cuando quisiera. ¿Es éste su caso? Si es así, ¿qué le impide trabajar? (INTERVIEWER LISTENS AND MARKS ALL THAT APPLY)

1 ☐ PRESTACIONES O BENEFICIOS ➔ GO TO EM5
2 ☐ PROBLEMAS DE SALUD ➔ GO TO EM5
3 ☐ NO TIENE INFORMACIÓN SOBRE RECURSOS LABORALES ➔ GO TO EM5
4 ☐ OTRAS PERSONAS LE HAN ACONSEJADO NO HACERLO ➔ GO TO EM5
5 ☐ NECESITA CAPACITACIÓN O EDUCACIÓN ➔ GO TO EM5
6 ☐ ESTÁ BUSCANDO TRABAJO PERO NO ENCUENTRA ➔ GO TO EM5
7 ☐ PROBLEMAS CON EMPLEO ANTERIOR ➔ GO TO EM5
8 ☐ TRANSPORTE ➔ GO TO EM5
9 ☐ CUIDADO DE LOS HIJOS ➔ GO TO EM5
10 ☐ OTRO (_____) ➔ GO TO EM5
11 ☐ NADA SE LO IMPIDE ➔ GO TO EM5

-1 ☐ NO SABE ➔ GO TO EM5
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO EM5
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO EM5

EM4. A veces a uno le gustaría tener un trabajo por el cual le paguen, pero le parece que algo se lo impide. ¿Es éste su caso? Si es así, ¿qué le impide querer trabajar? (INTERVIEWER LISTENS AND MARKS ALL THAT APPLY)

1 ☐ PRESTACIONES O BENEFICIOS ➔ GO TO THE ABOUT YOU SECTION
2 ☐ PROBLEMAS DE SALUD ➔ GO TO THE ABOUT YOU SECTION
3 ☐ NO TIENE INFORMACIÓN SOBRE RECURSOS LABORALES ➔ GO TO THE ABOUT YOU SECTION
4 ☐ OTRAS PERSONAS LE HAN ACONSEJADO NO HACERLO ➔ GO TO THE ABOUT YOU SECTION
5 ☐ NECESITA CAPACITACIÓN O EDUCACIÓN ➔ GO TO THE ABOUT YOU SECTION
6 ☐ ESTÁ BUSCANDO TRABAJO PERO NO ENCUENTRA ➔ GO TO THE ABOUT YOU SECTION
7 ☐ PROBLEMAS CON EMPLEO ANTERIOR ➔ GO TO THE ABOUT YOU SECTION
8 ☐ TRANSPORTE ➔ GO TO THE ABOUT YOU SECTION
8 ☐ CUIDADO DE LOS HIJOS ➔ GO TO THE ABOUT YOU SECTION
10 ☐ OTRO (_____) ➔ GO TO THE ABOUT YOU SECTION
11 ☐ NADA, NO QUIERE TRABAJAR ➔ GO TO THE ABOUT YOU SECTION
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM5. ¿Ha pedido ayuda para encontrar trabajo por el cual le paguen?

1 ☐ SÍ ➔ GO TO EM7
1 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM6. ¿Sabe que puede recibir ayuda para encontrar un trabajo por el cual le paguen?

1 ☐ SÍ ➔ GO TO THE ABOUT YOU SECTION
2 ☐ NO ➔ GO TO THE ABOUT YOU SECTION
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM7. La ayuda para encontrar trabajo puede incluir ayuda para encontrar un lugar en donde trabajar o ayuda para recibir la capacitación que necesita para trabajar. ¿Le pagan a alguien para que le ayude a encontrar trabajo?

1 ☐ SÍ ➔ GO TO EM8
2 ☐ NO ➔ GO TO THE ABOUT YOU SECTION
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM8. ¿Está recibiendo toda la ayuda que necesita para encontrar un trabajo?

1 ☐ SÍ ➔ GO TO THE ABOUT YOU SECTION

2 ☐ NO ➔ GO TO THE ABOUT YOU SECTION
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM9. ¿Quién le ayudó a encontrar el trabajo que tiene ahora? [MARK ALL THAT APPLY]

1 ☐ PERSONAL LABORAL / PERSONAL VOCACIONAL / ENTRENADOR LABORAL
2 ☐ ENCARGADO DE CASO
3 ☐ OTROS PROVEEDORES A QUIENES SE LES PAGA
4 ☐ OTROS SERVICIOS DE ORIENTACIÓN LABORAL
5 ☐ FAMILIARES O AMIGOS
6 ☐ ANUNCIO PUBLICITARIO
7 ☐ TRABAJA POR SU CUENTA ➔ GO TO EM11
8 ☐ OTRO (_____)
9 ☐ NADIE LE AYUDÓ. LO ENCONTRÓ SOLO(A) ➔ GO TO EM11
-1 ☐ NO SABE ➔ GO TO EM11
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO EM11
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO EM11

EM10. ¿Contribuyó a escoger el trabajo que tiene ahora?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM11. A veces uno necesita ayuda de los demás para hacer su trabajo. Por ejemplo, puede necesitar ayuda para llegar al trabajo o para ir de un lugar a otro en el sitio donde trabaja, para hacer el trabajo o para llevarse bien con los otros empleados. ¿Le pagan a alguien para ayudarle en el trabajo que tiene ahora?

1 ☐ SÍ
2 ☐ NO ➔ GO TO THE ABOUT YOU SECTION]
-1 ☐ NO SABE ➔ GO TO THE ABOUT YOU SECTION
-2 ☐ SE NEGÓ A CONTESTAR ➔ GO TO THE ABOUT YOU SECTION
-3 ☐ RESPUESTA POCO CLARA ➔ GO TO THE ABOUT YOU SECTION

EM12. ¿Cómo llama a esta persona? ¿Entrenador laboral (*job coach* en inglés), proveedor de apoyo para personas en la misma situación (*peer support* en inglés), asistente personal o alguna otra cosa?

_____

[USE THIS TERM WHEREVER IT SAYS {*job coach*} BELOW.]

EM13. ¿Contrató usted mismo a su {*entrenador laboral*}?

1 ☐ SÍ ➔ GO TO THE ABOUT YOU SECTION
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM14. ¿Su {*entrenador laboral*} está con usted todo el tiempo que usted está trabajando?

1 ☐ SÍ

2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM15. ¿Con qué frecuencia su {*entrenador laboral*} le da toda la ayuda que usted necesita? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Su {*entrenador laboral*} le da toda la ayuda que usted necesita? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM16. ¿Con qué frecuencia su {*entrenador laboral*} es amable y educado con usted? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Su {*entrenador laboral*} es amable y educado con usted? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM17. ¿Con qué frecuencia su {*entrenador laboral*} le explica cosas de una manera fácil de entender? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Su {*entrenador laboral*} le explica cosas de una manera fácil de entender? ¿Diría que…?

1 ☐ En general, sí, o

2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM18. ¿Con qué frecuencia su {*entrenador laboral*} lo(a) escucha con atención? ¿Diría que…?

1 ☐ Nunca,
2 ☐ A veces,
3 ☐ Casi siempre, o
4 ☐ Siempre?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

Versión Alternativa: ¿Su {*entrenador laboral*} lo(a) escucha con atención? ¿Diría que…?

1 ☐ En general, sí, o
2 ☐ En general, no?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM19. ¿Su {*entrenador laboral*} lo(a) anima a hacer cosas sin ayuda si usted puede hacerlas?

1 ☐ SÍ
2 ☐ NO
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM20. ¿Usando un número del 0 al 10, el 0 siendo la peor ayuda que recibe del{*entrenador laboral*} posible y el 10 es la mejor ayuda que recibe del {*entrenador laboral*} posible, ¿qué número usaría para calificar la ayuda que recibe del {*entrenador laboral*}?

___ 0 a 10

-1 ☐ DON'T KNOW ➔ GO TO ALTERNATE VERSION
-2 ☐ REFUSED
-3 ☐ UNCLEAR RESPONSE ➔ GO TO ALTERNATE VERSION

Versión Alternativa: ¿Cómo calificaría la ayuda que recibe del {*entrenador laboral*}? ¿Diría que es…?

1 ☐ Excelente,
2 ☐ Muy buena,
3 ☐ Buena,
4 ☐ Regular, o
5 ☐ Mala?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA

EM21. ¿Les recomendaría a sus familiares y amigos el {*entrenador laboral*} que le ayuda a usted si ellos necesitaran {*término específico del encuestado para "servicios que presta un entrenador laboral"*}? ¿Diría que les recomendaría el {*entrenador laboral}*?

1 ☐ Definitivamente no,
2 ☐ Probablemente no,
3 ☐ Probablemente sí, o
4 ☐ Definitivamente sí?
-1 ☐ NO SABE
-2 ☐ SE NEGÓ A CONTESTAR
-3 ☐ RESPUESTA POCO CLARA