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 #: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure
Measure Steward: American Health Care Association/National Center for Assisted Living
Brief Description of Measure: The measure calculates the percentage of family or designated responsible party for assisted living (AL) residents. This consumer reported outcome measure is based on the CoreQ: AL Family Satisfaction questionnaire that has three items.
Developer Rationale: Collecting satisfaction information from Assisted Living (AL) residents and family members 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 long-term care 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). We have developed three SNF based CoreQ measures, and these are NQF endorsed. But no equivalent instrument exists for AL.

Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Center for Excellence in Assisted Living (CEAL) which has developed a measure of person-centeredness of assisted living with UNC, 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 long-term care 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 long-term care 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: AL Family Satisfaction questionnaire and measure can strategically help AL 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 AL patients is tenable. A review of the literature on satisfaction surveys in long-term care facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many 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 (with 100% as a maximum score).

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: AL Family Satisfaction Measure has considerable relevance in establishing benchmarking scores and comparison scores. AHCA/NCAL developed three skilled nursing facility (SNF) based CoreQ measures: CoreQ: Long-Stay Family Satisfaction Measure, CoreQ: Long-Stay Resident Satisfaction Measure, and CoreQ: Short-Stay Discharge Measure. All three of these measures received NQF endorsement in 2016. In addition to the CoreQ Family Satisfaction Measure, AHCA/NCAL is submitting a CoreQ: Resident Satisfaction Measure. With these five satisfaction measures, it enables providers to measure satisfaction across the long term care continuum with valid and reliable measures.

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. Some assisted living communities have implemented QAPI in their organizations.

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 Core Q: AL 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 AL residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party for AL residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: AL Family Satisfaction questionnaire.

Denominator Statement: The target population is family or designated responsible party members of a resident residing in the facility for at least two weeks. The denominator includes all of the individuals in the target population who respond to the CoreQ: AL Family Satisfaction questionnaire within the two month time window who do not meet the exclusion criteria

Denominator Exclusions: Exclusions made at the time of sample selection are the following: (1) Court-appointed guardian; (2) family of residents receiving hospice; (3) Family members who reside in another country and (4) family of residents who have lived in the AL facility for less than two weeks.

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) and b) surveys that have more than one questionnaire item missing.

Measure Type: Outcome: PRO-PM
Data Source: Instrument-Based Data
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 outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- The developer provides a logic model outlining the relationship between the outcome of Assisted Living (AL) resident satisfaction and drivers such as staff competency, concern, and responsiveness of management.
- The developer provides eleven sources of evidence pertaining to value and meaningfulness of the outcome of satisfaction. Evidence includes two systematic reviews and other research specific to patient-clinician relationships and healthcare outcomes.
- The developer provides data demonstrating structure and process drivers associated with outcome of customer satisfaction. Studies from both assisted living and nursing homes are included.

Question for the Committee:

- *Is there at least one thing that the provider or facility can do to achieve a change in the measure results?*
- *Does the target population value the measured outcome and find it meaningful? Does the Committee feel the evidence provided support this?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

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 results from testing based on data from 463 AL facilities from multiple states. The performance scores show great variation between facilities demonstrating an opportunity for improvement.

CoreQ: AL Family Satisfaction measure (expressed in percent based on the CoreQ 0 – 100 scale)				
Minimum	**25th percentile**	**50th percentile**	**75th percentile**	**Maximum**
22%	57%	67%	75%	95%

Disparities

- The measure was not risk adjusted by sociodemographic status due to no statistically significant differences in the score between SDS categories of race, education and age. See results in Table 2b3.4b.c.

- Additional literature is provided on disparities in nursing facilities suggesting that social economic status differences are related to inter-facilities differences and not to intra-facility differences in care.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does the Committee agree that the disparities evidence at the SNF level is applicable to assisted living level?*

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

- My comments from 3420 are relevant to 3422 - in I am having a hard time connecting satisfaction being an indicator of high quality.
- meets - same comments as 3420
- Evidence is adequate.
- The evidence appears outdated. Also, the reference to the LTC Survey is questionable as I believe NQF no longer endorses that survey.
- The developers don't directly address importance to family members of Assisted Living residents. The evidence presented pertains to patient satisfaction. The focus groups conducted are described as including residents and family members, but the number of family members included is not reported.
- This is a patient reported outcome on patient experience with evidence provided regarding the association of experience with a number of outcomes. There are existing surveys for populations in skilled long term care and this is now being applied to assisted living where it will be a new measure.
- excellent relationship of data to outcome. Direct.

1b. Performance Gap

- My biggest concern here is the differences between the 25th percentile and the 75th percentile. It does not seem that there is a very big gap between those ranges.
- meets - same comment as 3420
- will need to address multiple languages
- the literature referenced is quit dated, but understand there is little research in the AL industry; some concern comparing AL with nursing home industry
- Gap is sufficient to make measure somewhat useful to residents/families facing choice of facilities.
- The value of evaluating satisfaction in Assisted Living is certainly important. Any objective measures to aid in consumer selection would be good.
- Yes, the range and IQR of measure results shown through testing show moderate performance gap.
- There is variation in the results suggesting room for improvement with specific questions that could help inform areas to improve.
- wide variation in scores indicate opportunities for improvement using this process

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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 – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators:

- David Cella, PhD, co-chair
- Karen Joynt Maddox, MD, MPH, co-chair
- Maybeth Farquhar, PhD, MSN, RN
- Paul Gerrard, BS, MD
- Eugene Nuccio, PhD

Evaluation of Reliability and Validity:

Evaluation A, Evaluation B, Evaluation C, Evaluation D, Evaluation E

Additional Information regarding Scientific Acceptability Evaluation:

Additional co-chair evaluations were needed for this measure. One of the original reviewers found both the reliability and validity testing to be insufficient. Both co-chairs gave the measure moderate ratings for reliability and validity.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *Five members of the Scientific Methods Panel reviewed the measure, four of the reviewers are satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Five members of the Scientific Methods Panel reviewed the measure, three of the reviewers are satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

Committee pre-evaluation comments

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

2a1. Reliability-Specifications

- meets - same comment as 342
- Specifications seem clear.
- The data elements were clearly defined. My question is whether the demographics of the population tested reflect the demographics of the broader assisted living population (90% white, 50% with higher education). For example, were the facilities tested more private pay, were some subsidized, etc?
- good understanding of measures by participants, good inter-rater reliability

2a2. Reliability – Testing

- I agree with the moderate rating for reliability considering the scientific panel's recommendations.
- meets, but a few comments (in addition to those in 3420)
- one geographic area for test with a 97% response rate
- No concerns.
- Reliability was rated moderate. Sample size was small and perhaps not adequate to evaluate reliability.
- Do not disagree with the overall assessment of the methods panel that the measure is reliable. The one dissenting methods panelist raised questions about methods, but did not dispute results which showed adequate reliability.
- No
- No
- Reliability: Although the patient level reliability was tested appropriately (they did internal consistency and test-retest reliability), the facility level testing was not. What is called 'two stage signal-to-noise' analysis is really bootstrapping within facility across multiple patient samples. What is needed is intraclass correlation coefficients that look at the ratio of between vs. between plus within facility variance. The later includes the across patients within the facility (or error) variance. Some of the reviewers from the Scientific Methods Panel (I wasn't one!) pointed that out.

2b1. Validity -Testing

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)

2b4. Meaningful Differences

- meets
- face and content validity - strong correlations
- missing responses low and addressed sound methodology for inclusion
- not sure it makes sense to exclude family members of hospice patients (understood that with patients)
- No concerns.
- Do not disagree with the overall assessment of the methods panel that the measure is valid. Two dissenting methods panelists raised questions about methods, but did not dispute results which showed adequate validity.
- No concerns
- unclear how missing data might affect the scores but this does not appear to be a major factor
- Validity: They provide some evidence for face validity and 'readability' of the questionnaire (Flesch-Kinkaid assessment), and some association with other indicators of quality (e.g. High staff turnover, crowding, etc.), but the associations are weak. They said they did exploratory factor analysis, but with 4 items, that really isn't appropriate either. That kind of analysis is intended to identify multiple dimensions within a construct (how many could you find with 4 items?). They should have at least done principal components analysis and given us some statistics associated with it (e.g. Percent variance explained by the first factor). It would have been far more convincing to have compared their measure with CAHPS for SNFs for example. The distribution of facility level mean scores provides us no evidence for meaningful differences; providing the standard error bars would have helped. For example, for Table 2b4.2, scores are favorably skewed. it would not be surprising if the highest and lowest scoring facilities have the lowest sample sizes and therefore most likely the highest within facility (error) variance.
- The samples for both measures are predominantly white and very well educated. Their findings may not generalize to a more diverse patient/facility population.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)

- no risk adjustment; well addressed in document
- No concerns.
- Measure is not risk adjusted, which seems appropriate given the testing results and what evidence can be extrapolated from nursing homes literature.
- My question is whether the demographics of the population tested reflect the demographics of the broader assisted living population (90% white, 50% with higher education). For example, were the facilities tested more private pay, were some subsidized, etc.? The results were segmented within this population for race and education and gender without variation noted and if it representative of the overall assisted living population in the majority of facilities, then this would likely apply.
- exclusions are reasonable, no risk adjustment. Low sampling of non-white populations

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.

- Measure is based on patient/family reported information in either paper or electronic format
- Measure is based on a 20 survey response sample
- No fees, licensing, or other requirements are associated with the measure

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

- Meets
- some extra time for caregivers to complete, but purely voluntary and may assist in engagement
- No concerns.
- Data elements not routinely generated, but originate from a very short survey. No major feasibility concerns.
- Identification of involved family members will be needed for the survey.
- requires a survey which is outside of usual care

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

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- Measure is currently used by the National Center for Assisted Living (NCAL) as part of their recognition program and has been picked up by seventeen national satisfaction vendors.
- This is a new measure and is not used in a public reporting program. The developer is working with states who require satisfaction measurement to promote the measure.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

N/A

Additional Feedback:
N/A

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**
RATIONALE:

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results
N/A new measure.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation
- 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: AL Family Satisfaction questionnaire or the measure that is calculated using this questionnaire.

Potential harms

- 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.

Additional Feedback:
N/A

***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

4a. Use
- Similar to my comments in 3420.

- Meets
- same comments as 3420
- Measure is new and adoption still underway. The prospect for future use looks promising; moreso than many new measures we review.
- It is not used for public reporting - once in the field it could be. Vendors have agreed to include this as an option and the survey content would help assisted living facilities in their improvement activities. Similar questions are being used in long term care facilities. There may be a history of family and patient involvement in the development of the earlier measures, I don't recall it being mentioned here
- Unclear

4b. Usability

- Similar to my comments in 3420, is this satisfaction information enough to guide our mission to higher quality and value care?
- Meets
- measure will be very usable; do wonder if there is a way to combine with 3422
- did not see details on how best to reach this group (esurvey, mail, telephone)"
- I doubt that this measure provides sufficient information to a facility to use the results for internal QI purposes. For use by purchasers and prospective residents and their families it will have to be publicly reported. I trust that will happen soon if measure is endorsed.

- Evidence review supports quite a few structural, process or organizational drivers of satisfaction that AL facilities can use in trying to make improvements. Most of that evidence seems focused on patient rather than family satisfaction, however. Would want to see usability reevaluated in the future once the measure is more widely adopted.
- The results can be used to target improvement. Patient perception of care is related to some of the outcomes noted in the testing and in the literature. The unintended consequences could be if the testing was done in higher income facilities and is now used in facilities where housing is subsidized, it's possible there may be disparities not recognized in the initial sample. Perhaps this could be assessed over time.
- sufficiently granular to point facilities towards interventions to improve satisfaction scores.

Criterion 5: Related and Competing Measures

Related or competing measures

Related Measures

- 2614 : CoreQ: Short Stay Discharge Measure
- 2615 : CoreQ: Long-Stay Resident Measure
- 2616 : CoreQ: Long-Stay Family Measure
- 3420 : CoreQ: AL Resident Satisfaction Measure

Harmonization

- All of the related measures have been developed by the same developer and are harmonized to the extent possible.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Evaluation A for Scientific Acceptability

Measure Number: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure

Scientific Acceptability: 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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   ***TIPS**: Consider the following: 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?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2
   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☐High (go to Question #6)

   ☒Moderate (go to Question #6)

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☒Yes (go to Question #7)

   ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)
☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:
    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?
REFERENCE: Testing attachment, section 2b2.
TIPS*: Consider the following: 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)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*
☐Yes (please explain below then go to Question #13)
☒No (go to Question #13)
☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?
TIPS*: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? 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, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
☐Yes (please explain below then go to Question #14)
☐No (go to Question #14)
☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.
☐Yes (please explain below then go to Question #15)
☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☒Yes (please explain below then go to Question #17)
☐No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

The developers report that exploratory factor analysis was done and that it supports a unidimensional construct. With only 3 items, this is an odd way of reporting their results, since it is mathematically impossible for a number of factor analysis computation techniques to converge to a solution with more than one factor with so few items, and if even a two factor model could be developed, it would be tantamount to saying that one of the three items is not like the other two rather than showing a multifactor solution. However, this does shows that a large proportion of variance is explained by a single factor. Additionally, the results of the survey have criterion validity in their correlation with other measures of the facility (table 2b.1.3f), I think that an important type of validity is shown, and this measure seems to have at least some evidence of criterion validity, indicating that the data it provides has some degree of inferential quality even in the absence of knowing the results of the factor analysis.

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)

☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☒Yes (go to Question #22)
☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
***REFERENCE:** Testing attachment, section 2b1.*
***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☒Yes (go to Question #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
☒Moderate (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*
☐Yes (go to Question #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of <u>all</u> testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there <u>are</u> threats to validity and/or threats to validity were <u>not assessed</u>]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level <u>is required</u>; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Evaluation B for Scientific Acceptability

Measure Number: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure

Scientific Acceptability: 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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

27. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: 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?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

28. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

In my opinion, percent agreement has meaning but it has too many weaknesses to be the sole indicator of reliability. Should use an additional method to verify that results were not attributed to chance.

29. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

30. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

31. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

32. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

33. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☒No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

Only assessed percent agreement.

34. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☒Insufficient (go to Question #9)

35. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☒No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

Only used face validity.

OVERALL RELIABILITY RATING

36. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☒Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

See above comments.

VALIDITY

Assessment of Threats to Validity

37. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

38. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: 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)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

39. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? 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, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

40. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

No like comparisons to other existing satisfaction surveys.

41. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)
☐Not applicable (go to Question #16)

42. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

43. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☐Yes (go to Question #18)
☒No (please explain below, then skip Questions #18-23 and go to Question #24)
Only used face validity.

44. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☐Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

45. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☐Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

46. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)
☐Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

47. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☐Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

48. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

REFERENCE: *Testing attachment, section 2b1.*

TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

49. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

50. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☒No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

Used pilot questionnaire (18 items) to assess validity of the three-item questionnaire. No info on the 18-item questionnaire to indicate it was valid. Used the literature and focus groups for establishing face validity.

51. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

52. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☒Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Empirical Validity or predictive validity is the relationship between test scores and some criterion of performance obtained, in this case satisfaction. I am concerned that the developer did not consider the testing that was done on the instruments did not rise to the level required for empirical testing. Perhaps correlation of the instrument with an established satisfaction survey.

Evaluation C for Scientific Acceptability

Measure Number: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure

Scientific Acceptability: 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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

53. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: 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?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

54. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

55. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

56. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

I believe that the two-level signal to noise test described of which the results were reported was actually an ICC, where the level of analysis was the facility.

57. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

58. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

59. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
REFERENCE: Testing attachment, section 2a2.2
***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☒Yes (go to Question #8)
☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

60. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
REFERENCE: Testing attachment, section 2a2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
☐Insufficient (go to Question #9)

61. Was **empirical VALIDITY testing** of patient-level data conducted?
REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

62. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:
☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

63. Were potential threats to validity that are relevant to the measure empirically assessed ()?
REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

64. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: 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)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

65. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? 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, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

66. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

67. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)
☐Not applicable (go to Question #16)

68. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☒Yes (please explain below then go to Question #17)
☐No (go to Question #17)

Imputation introduces a certain amount of error; it was unclear to me whether this was necessary to do, or whether missing data could just be considered missing.

Assessment of Measure Testing

69. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

70. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

71. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

72. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

73. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

74. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

75. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

76. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

77. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

78. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐ High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒ Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐ Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐ Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Evaluation D for Scientific Acceptability

Measure Number: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure

Scientific Acceptability: 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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

79. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: 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?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

80. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

81. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

82. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

83. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

84. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

85. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

86. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

87. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

88. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

89. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

90. Analysis of potential threats to validity: Any concerns with measure exclusions?
REFERENCE: Testing attachment, section 2b2.
TIPS*: Consider the following: 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)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*
☐Yes (please explain below then go to Question #13)
☒No (go to Question #13)
☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

91. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?
TIPS*: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? 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, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
☒Yes (please explain below then go to Question #14)
☐No (go to Question #14)
☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

They evaluated some possible risk adjusters but then provided no risk adjustment. Results were inconclusive

92. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.
☒Yes (please explain below then go to Question #15)
☐No (go to Question #15)

They studied differences by site, but without statistical testing or interpretive anchors.

93. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☒No (go to Question #16)
☐Not applicable (go to Question #16)

94. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☒Yes (please explain below then go to Question #17)
☐No (go to Question #17)

I'm troubled by the use of responding participants as the numerator. How will this migrate to application of an entire patient population? How will non-responders be handled? Mean imputation is limited

Assessment of Measure Testing

95. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

96. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

97. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

98. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

99. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

100. Was the method described and appropriate for assessing the accuracy of ALL critical data elements?

NOTE that data element validation from the literature is acceptable.

REFERENCE: *Testing attachment, section 2b1.*

TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

101. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

102. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

103. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

104. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐ High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒ Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐ Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐ Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Evaluation E for Scientific Acceptability

Measure Number: 3422
Measure Title: CoreQ: AL Family Satisfaction Measure

Scientific Acceptability: 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.

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

105. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE: *NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS: *Consider the following: 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?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

106. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

107. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

108. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

109. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

110. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☒Yes (go to Question #7)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

111. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #8)
☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

112. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
REFERENCE: Testing attachment, section 2a2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
☐Insufficient (go to Question #9)

113. Was **empirical VALIDITY testing** of patient-level data conducted?
REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.
☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

114. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:
☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

115. Were potential threats to validity that are relevant to the measure empirically assessed ()?
REFERENCE: Testing attachment, section 2b2-2b6
***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

116. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS*: Consider the following: 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)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

Relationship between Table 2b1.3.c: Respondent's Understanding of Response Scale and Part 2 A where "Each family member was asked to rate on a scale of 1 to 10…" appears to be inconsistent.

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

117. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

TIPS*: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? 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, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☒Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

This is based on their response to question 2b3.1. However, their response to question 1.8 indicated that they investigated four possible socio-demographic risk factors—including race that should not be used as a risk factor.

118. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

Table 2b4.2a shows a range of scores for different facilities (not clear). No test of statistical difference among the facilities was offered based on the measure score. Table 2b5.2 offers differences between individual correlated items to the measure, but not on the measure itself. I was hesitant to evaluate this as a "yes".

119. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☒No (go to Question #16)
☐Not applicable (go to Question #16)

120. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

121. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*
☒Yes (go to Question #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)
One way to address validity of self-report satisfaction items is to look at the predictive validity of these items. That is, do responses to these items track in the same direction as other results or items from other instruments. The answer to this is either "partially" (i.e., most reported results tracked in the expected direction but some did not and there was no explanation for these discrepancies offered by the Developer) or "maybe" (i.e., instrumentation methodology was described but there was no explicit evidence that the methodology was followed).

122. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

123. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*
☒Yes (go to Question #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

124. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

125. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)
☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

126. Was the method described and appropriate for assessing the accuracy of ALL critical data elements?

NOTE that data element validation from the literature is acceptable.
***REFERENCE:** Testing attachment, section 2b1.*
***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

127. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

128. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

129. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS: *Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

130. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☒Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Needs more information provided by developer regarding validity testing as described above. The measure developer notes that the measure is not risk adjusted—in a manner similar to CAHPS results. Simple demographics (e.g., male/female; age groups) could be used to stratify the results for quality improvement purposes.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: *CoreQ: AL Family Satisfaction 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: Click here to enter a date

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ 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. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

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

☐ Process: Click here to name what is being measured

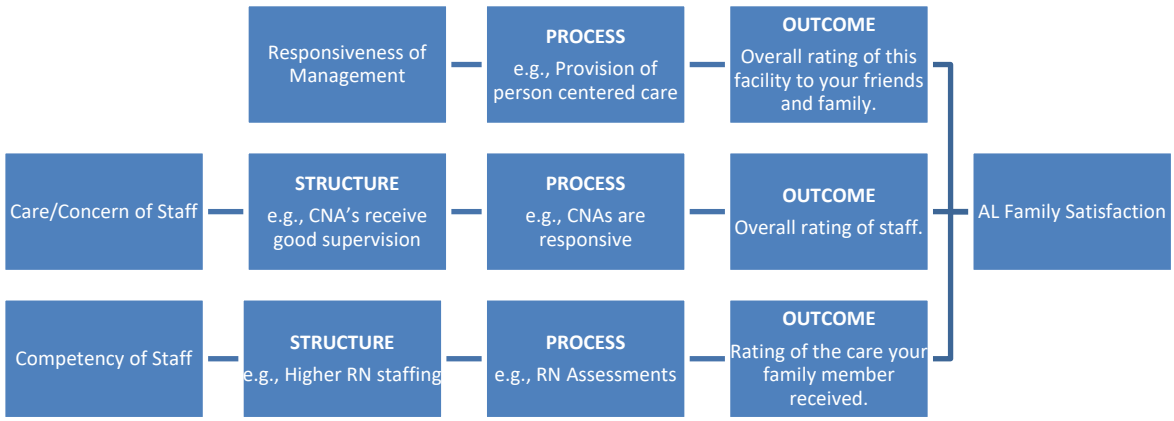
☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Family satisfaction can be looked at as the outcome for a number of structures and processes within Assisted Living (AL). 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.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** 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 AL facilities. A 2013 systematic review of studies on the effect of person-centered initiatives in long-term care 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 AL 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 long-term care residents with dementia and higher satisfaction with care (Van Uden et al., 2013). Improvements in a long-term care food delivery system also were associated with higher overall satisfaction and improved resident health (Crogan et al., 2013). The advantage of the CoreQ: AL Family Satisfaction questionnaire is it is broad enough to capture family's 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 Core Q: AL 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 AL facilities in

the Pittsburgh region. 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 Core Q: AL 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.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

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. We include studies from both AL and nursing homes. The nursing home studies are likely generalizable to AL

Table 1a.2.1: The structure and process drivers associated with AL resident satisfaction.

Authors	Structure or Process and Driver of AL Family Satisfaction	Summary Statement showing structures, processes, interventions and services and influence AL resident 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 Staff	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	Summary Statement showing structures, processes, interventions and services and influence AL resident 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	CNAs 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	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.

		council have higher resident satisfaction.	

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.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses

explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the 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 sub criterion 1b).

Brief Measure Information

NQF #: 3422
Corresponding Measures:
De.2. **Measure Title:** CoreQ: AL Family Satisfaction Measure
Co.1.1. **Measure Steward:** American Health Care Association/National Center for Assisted Living
De.3. **Brief Description of Measure:** The measure calculates the percentage of family or designated responsible party for assisted living (AL) residents. This consumer reported outcome measure is based on the CoreQ: AL Family Satisfaction questionnaire that has three items.
1b.1. **Developer Rationale:** Collecting satisfaction information from Assisted Living (AL) residents and family members 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 long-term care 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). We have developed three SNF based CoreQ measures, and these are NQF endorsed. But no equivalent instrument exists for AL.

Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Center for Excellence in Assisted Living (CEAL) which has developed a measure of person-centeredness of assisted living with UNC, 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 long-term care 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 long-term care 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: AL Family Satisfaction questionnaire and measure can strategically help AL 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 AL patients is tenable. A review of the literature on satisfaction surveys in long-term care facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many 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 (with 100% as a maximum score).

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: AL Family Satisfaction Measure has considerable relevance in establishing benchmarking scores and comparison scores. AHCA/NCAL developed three skilled nursing facility (SNF) based CoreQ measures: CoreQ: Long-Stay Family Satisfaction Measure, CoreQ: Long-Stay Resident Satisfaction Measure, and CoreQ: Short-Stay Discharge Measure. All three of these measures received NQF endorsement in 2016. In addition to the CoreQ Family Satisfaction Measure, AHCA/NCAL is submitting a CoreQ: Resident Satisfaction Measure. With these five satisfaction measures, it enables providers to measure satisfaction across the long term care continuum with valid and reliable measures.

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. Some assisted living communities have implemented QAPI in their organizations.

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 Core Q: AL 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 AL residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party for AL residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: AL Family Satisfaction questionnaire.
S.6. **Denominator Statement:** The target population is family or designated responsible party members of a resident residing in the facility for at least two weeks. The denominator includes all of the individuals in the target population who respond to the CoreQ: AL Family Satisfaction questionnaire within the two month time window who do not meet the exclusion criteria.
S.8. **Denominator Exclusions:** Exclusions made at the time of sample selection are the following: (1) Court-appointed guardian; (2) family of residents receiving hospice; (3) Family members who reside in another country and (4) family of residents who have lived in the AL facility for less than two weeks.

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) and b) surveys that have more than one questionnaire item missing.

De.1. **Measure Type:** Outcome: PRO-PM
S.17. **Data Source:** Instrument-Based Data
S.20. **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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
CoreQ_AL_family_evidence_FINAL.docx
1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
Collecting satisfaction information from Assisted Living (AL) residents and family members 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 long-term care 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). We have developed three SNF based CoreQ measures, and these are NQF endorsed. But no equivalent instrument exists for AL.

Further supporting person-centered care and resident satisfaction are ongoing organizational change initiatives. These include: the Center for Excellence in Assisted Living (CEAL) which has developed a measure of person-centeredness of assisted living with UNC, 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 long-term care 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 long-term care 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: AL Family Satisfaction questionnaire and measure can strategically help AL 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 AL patients is tenable. A review of the literature on satisfaction surveys in long-term care facilities (Castle, 2007) concluded that substantial improvements in resident satisfaction could be made in many 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 (with 100% as a maximum score).

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: AL Family Satisfaction Measure has considerable relevance in establishing benchmarking scores and comparison scores. AHCA/NCAL developed three skilled nursing facility (SNF) based CoreQ measures: CoreQ: Long-Stay Family Satisfaction Measure, CoreQ: Long-Stay Resident Satisfaction Measure, and CoreQ: Short-Stay Discharge Measure. All three of these measures received NQF endorsement in 2016. In addition to the CoreQ Family Satisfaction Measure, AHCA/NCAL is submitting a CoreQ: Resident Satisfaction Measure. With these five satisfaction measures, it enables providers to measure satisfaction across the long term care continuum with valid and reliable measures.

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. Some assisted living communities have implemented QAPI in their organizations.

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 Core Q: AL 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 maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Multiple data sources were used for testing the measure. The largest and most recent data source included 463 AL facilities from multiple states across the US (i.e., Pitt Research Data). The data were collected during 2017 and included responses from 29,693 patients. This shows, on the 0 – 100 scale used for the CoreQ: AL Family Satisfaction measure (expressed in percent), the minimum score is 22, the 25th percentile is 57, the 50th percentile is 67 the 75th percentile is 75 and the maximum score is 95.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) 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. By race, Whites averaged a score of 86.7, Blacks 86.6 and Asians 86.7; there were no observations for Native Hawaiians or other Pacific Islanders, American Indian or Alaskan Natives (Table 2b4.4b.d in the Testing section). By highest education level, those with some high school but who did not graduate averaged 86.5, high school graduates averaged 86.9, those with some college or a 2-year degree averaged 86.7, those with a 4-year college degree averaged 86.3, and those with more than a 4-year college degree averaged 86.8 (Table 2b3.4b.c in the Testing section). By age group, residents younger than 65 years old averaged 86.5, those 65-74 averaged 86.9, those 75-84 averaged 85.7, and those older than 85 averaged 86.8. Furthermore, by gender, males averaged 86.5 and females averaged 86.7.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**
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). No equivalent work in AL facilities exists; therefore, the nursing facility work is referenced here.

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 & 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. We believe the same is true for AL facilities.

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.

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: CoreQ: AL Family Satisfaction Measure
Date of Submission: 1/5/2018
Type of Measure:

☒ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐Intermediate Clinical Outcome	☐ Cost/resource
☐ Process *(including Appropriate Use)*	☐ 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, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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 social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. 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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [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]

2b3. 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 social risk 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.

2b4. 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.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. 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. The degree of consensus and any areas of disagreement must be provided/discussed.

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.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: CoreQ: AL Family Satisfaction questionnaire	☒ other: CoreQ: AL Family Satisfaction questionnaire, Pilot CoreQ: AL Family Satisfaction questionnaire

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*).

1.3. What are the dates of the data used in testing? June, 2014-September, 2017

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.20***)	**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 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 four data sources, one of which had additional variables collected for a subset of respondents:

1. The Pilot CoreQ: AL Family Satisfaction Questionnaire was examined using responses from 1,521 Family members or resident representatives from a national sample of AL facilities (Data Source #1).
    a. In addition, Family-level sociodemographic (SDS) variables were examined using this same sample of 1,521 Family members or resident representatives (#1 above) in AL facilities across the US. (Data Source #1).
2. Validity testing of the Pilot CoreQ: AL Family Satisfaction Questionnaire was examined using responses from 100 Family members or resident representatives from the Pittsburgh area. (Data Source #2).
3. Core Q: AL Family measure was examined using 375 facilities and included responses from 13,095 Family members or resident representatives. These AL facilities were located in multiple states across the US. (Data Source #3).
4. In addition, the CoreQ: AL Family Satisfaction measure was examined along with other outcome measures using a national sample of 486 facilities (with 29,693 family members) [Data Source #4].

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)
Data Source 1	92	89	1,521
Data Source 2	86	83	100
Data Source 3	96	90	13,095
Data Source 4	75	71	29,693

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*)
Data was used from the CoreQ: AL Family Satisfaction Questionnaire. The questionnaire was administered to all eligible AL family (with the exclusions described in the Specification part of this application). The testing and analysis included:

1. The Pilot CoreQ: AL Family Satisfaction questionnaire was examined using responses from 1,521 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 1,521 family members (Data #1 above) in AL facilities across the US.
2. Validity testing of the Pilot Core Q: AL Family Satisfaction questionnaire was examined using responses from 100 family members from the Pittsburgh area. (Data #2)
3. CoreQ: AL Family Satisfaction questionnaire measure was examined using 375 facilities and included responses from 13,095 family members or resident representatives. These AL facilities were located in multiple states across the US. (Data #3)
   [Note: Data source #4 above was used for facility level analyses, and is not included in the resident level of analysis]

The descriptive characteristics of the family members are given in the following table that includes information from all the data used (the education level and race information comes only from the sample described above with 1,521 respondents, as this data was not collected for the other samples).

Table 1.6: Respondent Demographics

Demograhpics		Percent

Are you male or female?	Male	29%
	Female	71%
What year were you born?	Average	1948
What is the highest grade or level of school that you have completed?	Some HS	2%
	HS or GED	16%
	Some College/ 2yr Degree	28%
	4yr College Degree	22%
	>4yr College Degree	33%
What is your race?	White	91%
	Black	4%
	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: AL Family Satisfaction Measure. The first focused on testing (e.g., reliability, validity, and exclusions) of the CoreQ: AL Family Satisfaction Questionnaire. The first source of data (pilot data) was utilized in developing and choosing the items to be included in the CoreQ: AL Family Satisfaction Questionnaire. This included using a questionnaire with 18 items. Below we call this the Pilot CoreQ: AL Family Satisfaction 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: AL Family Satisfaction Questionnaire was developed, a second source of data was used to test the validity of the CoreQ: AL Family Satisfaction Measure (i.e., facility and summary score validity). This second data source is described above (i.e. 375 facilities including responses from 13,095 family members [Data #3, above]).

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

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: AL Family Satisfaction 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: AL Family Satisfaction questionnaire had responses from 100 family members; we re-administered the survey to 100 of these same family members, and 97% responded. 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: AL Family Satisfaction 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

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 Core Q: AL Family questionnaire sample. We also conducted two-level signal-to-noise analysis which identifies two sources of variability, those between ratees (facilities) and those for each ratee (respondents). No imputed values were used in the analysis and only AL facilities with 20 or more responses were included.

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: AL Family Satisfaction questionnaire items, and the response per item for both the pilot survey of 100 family members and the re-administered survey of 100 family members (i.e., 97 responses). 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: AL Family Satisfaction Questionnaire Responses from the Pilot and Re-administered Survey

Questionnaire Item	Response	Percent [Pilot Survey (N=100)]	Percent [Re-Administered Survey (N=97)]
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	3%	3%
	Average	11%	11%
	Good	16%	16%
	Very Good	31%	30%
	Excellent	39%	40%
2. Overall, how would you rate the staff?	Poor	3%	3%
	Average	13%	12%
	Good	15%	16%
	Very Good	28%	29%
	Excellent	41%	40%
3. How would you rate the care your family member received?	Poor	4%	4%
	Average	15%	15%
	Good	18%	19%
	Very Good	25%	26%
	Excellent	38%	36%

*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 Core Q: AL 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
1. In recommending this facility to your friends and family, how would you rate it overall?	97.9%
2. Overall, how would you rate the staff?	95.9%
3. How would you rate the care your family member received?	95.9%

(2) PERSON/QUESTIONNAIRE LEVEL

Table 2a2.3.c shows the CoreQ: AL Family Satisfaction 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 97 family members. The person-level responses in the pilot survey are not statistically significantly different 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, 96% 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 Agreement between Responses per Item for Pilot and Re-Administered Surveys

Questionnaire Item	Response	Percent Person-Level Agreement in Response for the Pilot Survey (N=100) vs. Re-Administered Survey (N=97)
1. In recommending this facility to your friends and family, how would you rate it overall?	Poor	100%
	Average	98%
	Good	98%
	Very Good	99%
	Excellent	99%
2. Overall, how would you rate the staff?	Poor	100%
	Average	98%
	Good	97%
	Very Good	99%
	Excellent	99%
3. How would you rate the care you receive?	Poor	98%
	Average	98%
	Good	96%
	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)	99.5%	98.5%
	Good (3), Very Good (4), or Excellent (5)	99%	98%

(3) MEASURE (FACILITY) LEVEL

After having performed the 10,000-repetition bootstrap, 15% of bootstrap repetition scores were within 1 percentage point of the score under the original pilot sample, 29% were within 3 percentage points, 42% were within 5 percentage points, and 79% were within 10 percentage points. For the two-level signal-to-noise analysis for CoreQ: AL Family R=0.82, indicating that 82% of facilities true score can be attributed to ratings from the respondents (AL families) and remaining 18% is due to noise and differences among respondents. This result exceeds what is generally considered a good reliability coefficient of 0.8 (Campbell et al., 2010).

Campbell, JA, Narayanan, A., Burford, B., Greco, MJ. Validation of a multi-source feedback tool for use in general practice. Education in Primary Care, 2010, 21, 165-179.

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 Core Q: AL Family questionnaire data elements were highly repeatable, with pilot and re-administered responses agreeing between 97% and 100% 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: AL Family Satisfaction 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 125.

2b1. VALIDITY TESTING
2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. 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: AL Family Satisfaction questionnaire, four 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: AL Family Satisfaction questionnaire (i.e., response scale).

The second source of data, was pilot data collected from a national sample of 1,521 family members. This data was used in choosing the items to be used in the CoreQ: AL Family Satisfaction questionnaire (i.e., questionnaire items). This data was also used in examining Family-level sociodemographic (SDS) variables.

The third source of data (collected from 375 facilities) was used examine the validity of the CoreQ: AL Family Satisfaction measure (i.e., facility and summary score validity). These family members / AL facilities were from multiple states across the U.S.

The fourth source of data (collected from 487 facilities described in Section 1.5) was used to examine the correlations between the CoreQ: AL Family Satisfaction measure scores and other quality metrics from the facilities.

Thus, the following sections describe this validity testing:
1. Validity Testing of the questionnaire format used in the CoreQ: AL Family Satisfaction questionnaire (using data source 1, from above);
2. Testing the items for the CoreQ: AL Family Satisfaction 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 (CoreQ: AL Family Satisfaction measure) (using data source 3, from above);
4. Validity testing for the CoreQ: AL Family Satisfaction measure (also using data source 1, from above and data source 4).

1. Validity Testing for the Questionnaire Format used in the CoreQ: AL Family Satisfaction Questionnaire

A. The face validity of the domains used in the CoreQ: AL Family Satisfaction 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 AL facilities in the Pittsburgh region.
C. The face validity of the Pilot CoreQ: AL Family Satisfaction questionnaire response scale was also examined. The respondents were family members (N=40) with residents in five AL 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).

Reference: 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: AL Family Satisfaction Questionnaire

The second series of validity testing was used to further identify items that should be included in the CoreQ: AL Family Satisfaction 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: AL Family Satisfaction questionnaire survey was administered consisting of 18 items (N= 1,521 family members). The testing consisted of:

A. Family members were asked to rate the 18 different satisfaction questions related to their experience in AL. This was conducted with a pilot group of 40 family members in focus groups.
B. The Pilot CoreQ: AL Family Satisfaction questionnaire items performance with respect to the distribution of the response scale and with respect to missing responses. (Using 1,521 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 1,521 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: AL Family Measure).

The Core Q: AL 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. 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: AL Family Measure.

A. To determine if the 3 items in the CoreQ: AL Family Satisfaction questionnaire were a reliable indicator of satisfaction, the correlation between these three items (the "CoreQ: AL Family Satisfaction Measure") and ALL of the items on the Pilot CoreQ instrument was conducted.

B. We performed additional validity testing of the facility-level CoreQ: AL Family measure by examining the correlations between the CoreQ: AL Family Satisfaction measure scores and several quality metrics from the AL facilities. If the CoreQ: AL Family Satisfaction 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: AL Family Satisfaction measure.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

1. Validity Testing for the Questionnaire Format used in the CoreQ: AL Family Satisfaction Questionnaire

A. The face validity of the domains used in the CoreQ: AL Family Satisfaction 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 2b1.3.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: AL Family Satisfaction questionnaire items.

Table 2b1.3.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: AL Family Satisfaction questionnaire are shown. Note, the food domain was ranked third – but was excluded from the CoreQ: AL Family Satisfaction measure based on: 1) additional analyses showing that it was highly correlated with the overall domain; 2) food was in many cases not actually experienced by family members; 3) it was included in the CoreQ: Resident Satisfaction Measure -- thus, it added little to this family measure.

Table 2b1.3.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 your family member received?)	2

C. The face validity of the pilot CoreQ: AL Family Satisfaction questionnaire response scale was also examined. Table 2b1.3.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 2b1.3.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: AL Family Satisfaction 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: AL Family Satisfaction 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 AL 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 AL. Specifically, "In recommending this facility to your friends and family, how would you rate it overall?" had an average score of 8.9; "Overall, how would you rate the staff?" had an average score of 9.4; and, "How would you rate the care your family member received?" had an average score of 9.2. This shows a very pervasive influence of the satisfaction items with the experience of AL care.

B. The pilot Core Q: AL Family questionnaire items all 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) were 10.62 for Factor 1 and 0.87 for Factor 2. 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.

3. *To determine if a Sub-Set of Items could Reliably be used to Produce an Overall Indicator of Satisfaction (The CoreQ: AL Family Satisfaction 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 2b1.3.d. 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 3 sets of correlations, however the analysis was conducted examining all possible correlations between items.

Table 2b1.3.d: CoreQ: AL Family Satisfaction Questionnaire Example Item Correlations

	Family

Highest Correlation	Q6-Q9 (.789)
Next highest Correlation	Q9-Q8 (.781)
Next highest Correlation	Q10-Q8 (.755)

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. Therefore, this analysis was also used to eliminate items. Note, table 2b1.3.e again provides 3 sets of correlations, however the analysis was conducted examining all possible correlations between items.

Table 2b1.3.e: Secondary Correlation Analysis of CoreQ: AL Family Satisfaction Questionnaire Items

	Family
Q1 + last satisfaction item ADD	Q10(.910) Q6 (.904) Q2 (.900)
Q1 + ADD ADD	Q9 + Q6 (.889) Q2 + Q6 (.887) Q10 + Q6 (.877)
Q1 + ADD ADD	Q9 + Q6 (.905) Q10 + Q9 (.899) Q6 + Q2 (.894)

Thus, using the correlation information and factor analysis 3 items representing the CoreQ: AL Family Satisfaction questionnaire were identified.

4. ***Validity Testing for the Core Q: AL 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: AL Family Satisfaction questionnaire.

A. The items were all scored according to the rules identified elsewhere. The same scoring was used in creating the 3 item CoreQ: AL Family Satisfaction questionnaire summary score and the satisfaction score using the Pilot Core Q: AL Family questionnaire. The correlation was identified as having a value of 0.91.

That is, the correlation score between the actual "CoreQ: AL Family Satisfaction 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: AL Family Satisfaction measure by measuring the correlations between the CoreQ: AL Family Satisfaction measure scores and several other quality metrics from AL providers (see Table 2b1.3.f) . CoreQ: AL Family Satisfaction measure is the percentage of

family members of residents 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: AL Family Satisfaction questionnaire. The summary score from the 3 CoreQ: AL Family Satisfaction 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.

Table 2b1.3.f: Correlations between CoreQ: AL Family Satisfaction Measure and Quality Indicators

Quality Indicator	Correlation with Satisfaction Summary Score
Hospitalization	-0.018787
Rehospitalization	-0.138338*
Off-label use of antipsychotic drugs	-0.23704*
LPN Turnover	-0.278406*
Aide Turnover	-0.184795*
Administration Turnover	-0.10511*
DON Turnover	-0.018422
All Staff Turnover	-0.203063*
Occupancy	-0.04181

*Statistically significant at p=<0.05

2b1.4. 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: AL Family Satisfaction Questionnaire
A. The literature review shows that domains used in the Pilot CoreQ: AL Family Satisfaction 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 members 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: AL Family Satisfaction 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 AL 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 CoreQ: AL Family Satisfaction measure)

A. Using the correlation information of the CoreQ: AL Family Satisfaction questionnaire (18 items) and the 3 items representing the CoreQ: AL Family Satisfaction 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 CoreQ: AL Family Satisfaction Measure

A. The correlation of the 3 item CoreQ: AL Family Satisfaction 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 0.91. That is, the correlation score between the actual "CoreQ: AL Family Satisfaction Satisfaction 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: AL Family Satisfaction measure score adequately represents the overall satisfaction of the facility. This testing indicates a high degree of criterion validity.

B. *Relationship with Quality Indicators*

The 9 quality indicators examined had a moderate level of correlation with the CoreQ: AL Family Satisfaction measure. These correlations range from 0.27 to 0.01. The CoreQ: AL Family Satisfaction measure is associated with all of the 9 quality indicators in the direction hypothesized (that is higher CoreQ scores are associated with better quality indicator scores). This testing indicates a moderate degree of construct validity and convergent validity.
As noted by Mor and associates (2003, p.41) when addressing quality of long-term care facilities, "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 long-term care facilities is consistently moderate with respect to the correlations identified. Thus, it is not surprising that "very high" levels of correlations were not identified. As described in the literature, some correlation was identified in the direction as expected, which is in support of validity of the CoreQ: Family Satisfaction Measure.

_____

2b2. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b3*

2b2.1. 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: AL Family Satisfaction 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 375 AL facilities that have used the CoreQ: AL Family Satisfaction measure. These facilities were included in multiple states across the US (this is data source 3, from above).

2b2.2. 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 AL for less than 2 weeks; (4) Respondents who have more than one 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 375 facilities (described elsewhere). The exclusions were tracked and from these facilities included <1% Family members of residents with hospice; and <1% 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.

2b2.3. 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 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). In addition, we excluded residents on hospice, which aligns with a majority of CMS CAHPS surveys.

2b3. 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 2b4.

2b3.1. 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

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

2b3.2. If an outcome or resource use component 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 AL facilities. Testing on this (for nursing homes) 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) (RTI International, 2003). 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.

Education may influence responses to the questions asked. That is, respondents with lower education levels may not appropriately interpret the items. To address this, our items were written and tested to very low Flesh-Kincaid levels. In testing, no differences in average item scores were identified based on education levels ($p<.05$) (Table2b3.4b.c). A t-test analysis was used to compare the CoreQ mean scores, adjusting for race (Table 2b3.4b.d). This analysis demonstrated the CoreQ: AL Family Satisfaction measure is not significantly different based on race. Based on these results, education level makeup of the respondents or the racial makeup of the respondents does not appear to be related to this measure. We included these background characteristics for two reasons. First, to examine if any responses were different based on these factors (in no case were the responses different). Second, to examine the representativeness of the samples (the samples examined were representative of national AL figures).

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). No equivalent work in AL facilities exists; therefore, the nursing facility work is referenced here.

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 & 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. We believe the same is true for AL facilities.

The CoreQ AL Family Satisfaction Measure has been administered by mail, in-person, and on-line. The individual responses and CoreQ measure for in-person and on-line were not significantly different from the responses by mail ($p<.05$). Indicating no risk-adjustment would be necessary to account for the mode of administration.

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.

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.

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.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?
Not Applicable

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?
Not Applicable

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk 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.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

Analyses used to examine SDS factors include: (1) the summary score for each of the 3 CoreQ: AL Family Satisfaction questionnaire items; (2) the summary score for the CoreQ: AL Family Satisfaction measure; and (3) the summary score from the CoreQ: AL Family Satisfaction questionnaire measure at the facility level.

(1) Summary Score for each of the 3 CoreQ: AL Family Satisfaction Questionnaire Items
The summary score for each of the 3 CoreQ: AL Family Satisfaction 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 (Tables 2b3.4b.a., 2b3.4b.b). These analyses show that the individual item scores used in the CoreQ: AL Family Satisfaction measure are not significantly different based on either education level or race.

Table 2b3.4b.a. Mean CoreQ: AL Family Satisfaction Distribution Item by Level of Education

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	2% (n=15)	4.00	3.95	3.80
High school graduate or GED	16% (n=138)	4.05	4.90	3.80
Some college or 2 year degree	28% (n=246)	4.01	4.05	3.85
4 year college graduate	22% (n=195)	4.10	4.10	4.05
More than 4 year college degree	33% (n=291)	4.04	4.05	3.95
Rank Correlation		0.0051	0.027	0.008

Rank Correlation of items with education: none significant at p=0.05

Table 2b3.4b.b. Mean CoreQ: AL Family Satisfaction Distribution Item by Race

What is your race?	Respondents	Q1	Q2	Q3
		Mean	Mean	Mean
White	91% (n=805)	4.03	4.02	3.88
Black or African-American	4% (n=35)	3.98	3.95	3.85
Asian	1% (n=9)	4.10	4.10	3.90
Native Hawaiian or other Pacific Islander	0% (n=0)	0	0	0
American Indian or Alaskan Native	0% (n=0)	0	0	0
Two-sample t-test*	1 vs. 2	1.95	1.32	1.90
	1 vs. 3	0.99	0.91	0.58
	2 vs. 3	0.72	1.63	0.98

*Differences not statistically significant at p=0.05

(2) Summary Score for the CoreQ: AL Family Satisfaction Measure

The summary score for each of the 3 CoreQ: AL Family Satisfaction 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 (Tables 2b3.4b.c. 2b3.4b.d). These analyses show that the CoreQ: AL Family Satisfaction 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 2b3.4b.c: Mean CoreQ: AL Family Satisfaction Distribution by Level of 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	2% (n=15)	3.60
High school graduate or GED	16% (n=138)	3.62
Some college or 2 year degree	28% (n=246)	3.59
4 year college graduate	22% (n=195)	3.63
More than 4 year college degree	33% (n=291)	3.61
Rank Correlation		0.033

Table 2b3.4b.d: Mean CoreQ: AL Family Satisfaction Distribution by Race

What is your race?	Respondents	Measure Score
		Mean
White	91% (n=805)	3.60
Black or African-American	4% (n=35)	3.61
Asian	1% (n=9)	3.61
Native Hawaiian or other Pacific Islander	0% (n=0)	0
American Indian or Alaskan Native	0% (n=0)	0
Two-sample t-test		p-value
	1 vs. 2	0.79
	1 vs. 3	0.81
	2 vs. 3	0.77

Differences not statistically significant at p=0.05

(1) **Summary score from the CoreQ: AL Family Satisfaction Measure (at the facility level).**
The summary score for each of the 3 CoreQ: AL Family Satisfaction 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. A t-test a**nalysis was** used to compare the mean scores (Tables 2b3.4b.c and2b3.4b.d). This analysis demonstrated the CoreQ: AL Family Satisfaction 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 2b3.4b.c: CoreQ: AL Family Satisfaction Score with and without stratification 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	2% (n=15)	86.5	86.4	n.s
High school graduate or GED	16% (n=138)	86.9	86.1	n.s
Some college or 2 year degree	28% (n=246)	86.7	86.5	n.s
4 year college graduate	22% (n=195)	86.3	86.5	n.s
More than 4 year college degree	33% (n=291)	86.8	86.6	n.s

N.S. = Not significant at p=0.05

Table 2b3.4b.d: CoreQ: AL Family Satisfaction Score with and without stratification for Race

What is your race?	Respondents	Measure Score (Mean)		
		Score with SDS Characteristic vs. Without Characteristic		
White	91% (n=805)	86.7	86.4	n.s
Black or African-American	4% (n=35)	86.6	86.3	n.s
Asian	1% (n=9)	86.7	86.7	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

2b3.5. 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 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable

2b3.9. Results of Risk Stratification Analysis:
Not Applicable

2b3.10. 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

2b3.11. 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

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. 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 two analyses to examine whether the CoreQ AL Family measure captured clinically/practically meaningful differences between providers. First, we produced a histogram of the scores for the providers in the Pilot CoreQ: AL Family Satisfaction Questionnaire sample (figure 2b4.2.a). Second, we calculated the means of a series of quality metrics at each quintile of the CoreQ AL Family measure scores, and also calculated the correlation coefficients overall (2b4.2.b). The quality metrics were Hospitalization, Rehospitalization, Antipsychotic drugs, LPN Turnover, Aide Turnover, Administration Turnover, DON Turnover, All Staff Turnover, and Occupancy (Table 2b4.1). They represent the same time period as that used for the CoreQ: AL Family Satisfaction measure data collection. The data source is Pitt Research Data. This includes AHCA /NCAL participants providing data on the LTC Trend Tracker[sm] system, national chains working with Pitt, and individual facilities across the country participating in satisfaction research with Pitt.

Table 2b4.1: Definition of Quality Metrics

Quality Metric	Definition
Hospitalization	Percent of residents with hospital admissions
Rehospitalization	Percent of residents admitted directly from a hospital / # of residents sent back to the hospital within the next 30 days
Off-label use of antipsychotic drugs	Percent of residents with use of off-label antipsychotic drugs/ #of residents in community
LPN Turnover	Percent of LPNs leaving AL facility during 2017
Aide Turnover	Percent of Aides leaving AL facility during 2017
Administration Turnover	Percent of administration staff leaving AL facility during 2017
DON Turnover	Percent of DONs leaving AL facility during 2017
All Staff Turnover	Percent of LPNs leaving AL facility during 2017
Occupancy	Percent of beds occupied by residents

2b4.2. 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 in figure 2b4.2.a shows the distribution of scores, ranging from 21% to 95%. Table 2b4.2.b summarizes the selection of quality metrics by quintile of the CoreQ: AL Family Satisfaction Measure.

Figure 2b4.2.a: The distribution of the CoreQ AL Family Measure Score

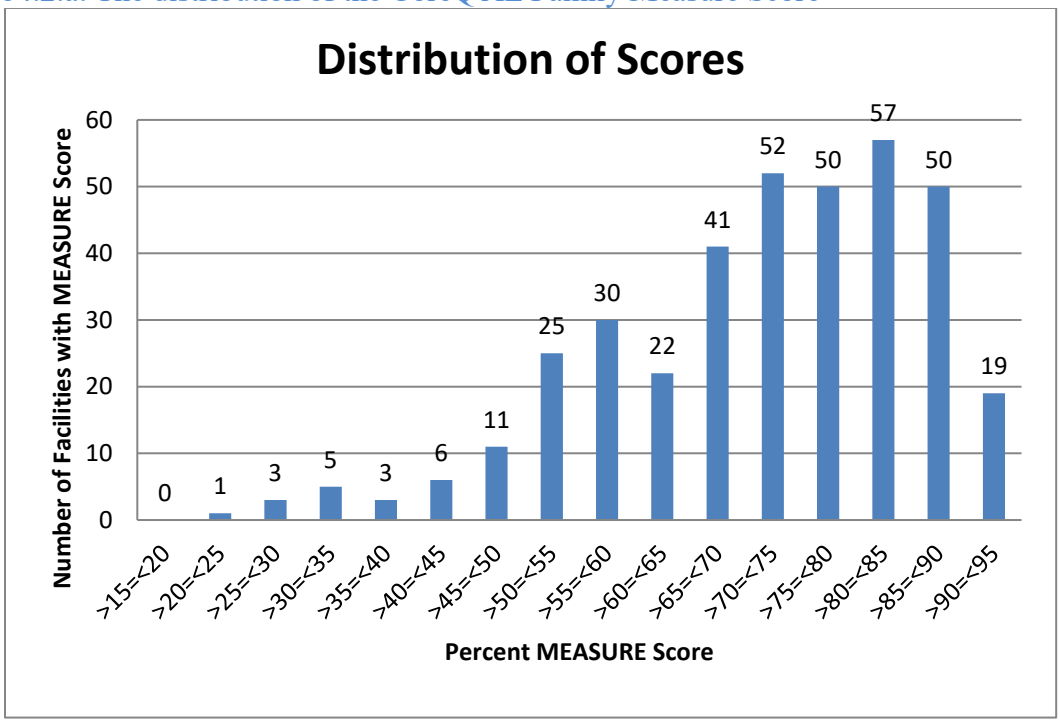


Table 2b5.2: Quality Metrics of the CoreQ: AL Family Satisfaction Measure

	Between CoreQ Score and Quality Metric	Mean Value	N (facilities)	Difference Between 5th and 1st Quintile
Range of CoreQ Scores	.22-.95			
Hospitalization	-0.018787	5.6%	487	1.9%
Rehospitalization	-0.138338	6.7%	487	2.2%
Antipsychotic drugs	-0.23704	6.6%	487	3.4%
LPN Turnover	-0.278406	12.0%	487	4.4%
Aide Turnover	-0.184795	20.0%	487	5.6%
Administration Turnover	-0.10511	37.0%	487	5.0%
DON Turnover	-0.018422	43.0%	487	5.5%
All Staff Turnover	-0.203063	20.0%	487	8.2%
Occupancy	-0.04181	87.0%	487	3.0%

2b4.3. 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 Core Q AL Family scores reflect practical and meaningful differences in quality between facilities. First, the histogram in Section 2b4.2 (figure 2b4.2.a) 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. Second, Table 2b5.2 shows the Core Q scores do indeed sort centers into those with high and low quality in other domains.

2b5. 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 social risk 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 social risk 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.***

2b5.1. 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

2b5.2. 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

2b5.3. 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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. 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: AL Family Satisfaction questionnaire. In calculating the CoreQ: AL Family Satisfaction 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: AL Family Satisfaction 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.

2b6.2. 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: AL Family Satisfaction questionnaire. In calculating the CoreQ: AL Family Satisfaction 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 13,095 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 2.63% (n=345) cases.
2. Overall, how would you rate the staff?
Missing responses occurred in 3.03% (n=397) cases.
3. How would you rate the care your family member received?
Missing responses occurred in 2.74% (n=360) cases.

Two (or more) missing responses occurred in 388 cases. Thus, the degree of missing data was very small (=2.96%). Imputation was used in 211 cases or 1.61% of respondents.

2b6.3. 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., Hospitalization, Rehospitalization, Antipsychotic drug use, LPN Turnover, Aide Turnover, Administration Turnover, DON Turnover, All Staff Turnover, and Occupancy) was unchanged. When the respondents were removed from the analyses, the average Summary Scores remained the same.

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 sub criteria 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. **Non-Condition Specific***(check all the areas that apply):* De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
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.)* coreq.org 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.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Attachment **Attachment:** Core_Q_–_Assisted_Living_Family_Satisfaction_Questionnaire.docx

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Family or other caregiver

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update 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) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The numerator assesses the number of family or designated responsible party for AL residents that are satisfied. Specifically, the numerator is the sum of the family or designated responsible party for AL residents that have an average satisfaction score of =>3 for the three questions on the CoreQ: AL Family Satisfaction questionnaire.

S.5. **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, time period for data collection, 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 (S.14).
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 for AL residents, they have up to 2 months to return the questionnaire. Only surveys returned within two months of the resident initially receiving the survey are included in the calculation.

The numerator includes all the family or designated responsible party members for AL residents that had an average response =>3 on the CoreQ: AL Family Satisfaction questionnaire.
We calculate the average satisfaction score for the individual family or designated responsible party member for AL residents in the following manner:
- Respondents within the appropriate time window and who do not meet the exclusions are identified.
- A numeric score is associated with each response scale option on the CoreQ: AL Family Satisfaction 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.

No risk-adjustment is used.

S.6. **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 the facility for at least two weeks. The denominator includes all of the individuals in the target population who respond to the CoreQ: AL Family Satisfaction questionnaire within the two month time window who do not meet the exclusion criteria.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The denominator includes all of the family or the designated responsible party members for residents that have been in the facility for at least two weeks or more regardless of payer status; who received the CoreQ: AL Family Satisfaction 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 facility records.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclusions made at the time of sample selection are the following: (1) Court-appointed guardian; (2) family of residents receiving hospice; (3) Family members who reside in another country and (4) family of residents who have lived in the AL facility for less than two weeks.

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) and b) surveys that have more than one questionnaire item missing.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*
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 facility for less than two weeks; and (4) all representatives 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.

No stratification is used.
Exclusions will be based on information from the facility health information system. Representatives of residents with the following criteria will be excluded:
(1) Residents on hospice. This is recorded in the facility health information system.
(2) Residents with court appointed legal guardian for all decisions will be identified from the facility health information system.
(3) Residents who have lived in the facility for less than two weeks days will be identified. This is recorded in the facility health information system.
(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.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Other (specify):
If other: Score is a percent and is not weighted.

S.13. **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.14. **Calculation Algorithm/Measure Logic** (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*
1. Identify the representatives of residents that have been residing in the facility for two weeks or more.
2. Take the representatives of residents that have been residing in the facility for >=two weeks and exclude the following:
a. Representatives of residents on hospice. This is recorded in the facility health information system.
b. Residents with Court appointed legal guardian for all decisions as identified from the facility health information system.
3. Exclude representatives of residents who reside in another country.
4. Administer the CoreQ: AL Family Satisfaction questionnaire 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: AL Family Satisfaction 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 AL 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: AL Family Satisfaction Measure which represents the percent of respondents with average scores of 3.0 or above.
a. CoreQ: AL Family Satisfaction Measure = ([number of respondents with an average score of =3.0] / [total number of valid responses])*100
9. No risk-adjustment is used.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
No sampling is used. All family representatives not meeting exclusions are to receive the survey. However, a minimum sample size of 20 and overall response rate of 30% is needed for the measure. This was based on analyses indicating that a minimum sample size of 20 and an overall response rate of 30% provided stable and representative scores.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
1. Identify the representatives of residents that have been residing in the facility for two weeks or more.
2. Take the representatives of residents that have been residing in the facility for >= two weeks and exclude the following:
a. Representatives of residents on hospice. This is recorded in the facility health information system.
b. Residents with Court appointed legal guardian for all decisions as identified from the facility health information system.
3. Exclude representatives of residents who reside in another country.
4. Exclude representatives of residents who died in the facility
5. Administer the CoreQ: AL Family Satisfaction questionnaire to family or designated responsible party members for AL residents.
6. Instruct representatives that they must respond to the survey within 2 months.
7. 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 facility.
8. Regardless of response rate, AL facilities 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.
All the questionnaires that are received (other than those that satisfy the exclusion criteria seen in section S.9) must be used in the calculations.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Instrument-Based Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
The collection instrument is the Core Q: AL Family questionnaire and exclusions are from the facility health information systems.

CoreQ: AL Family Satisfaction questionnaire
1. In recommending this facility to your friends and family, how would you rate it overall?
Poor Average Good Very Good Excellent
2. Overall, how would you rate the staff?
Poor Average Good Very Good Excellent
3. How would you rate the care your family member received?
Poor Average Good Very Good Excellent

Modes include in-person, mail, and online. Language of administration is English.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other
If other: Assisted Living

S.22. **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

2. Validity – See attached Measure Testing Submission Form

CoreQ__AL_Family_Testing_Final___Edited_1_31_18.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

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*) Update this field for **maintenance of endorsement**.
Patient/family reported information (may be electronic or paper)

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.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
Electronic sources are not used as this is a satisfaction survey

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. Please also complete and attach the NQF Feasibility Score Card.**
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. **Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Since the Core Q: AL 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 AL facilities were able to achieve the 20 survey response goal identified above. We identified that a majority of AL 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: AL Family Satisfaction measure development, existing satisfaction vendors were contacted (including NRC/MyInnerView, Pinnacle, Providigm, and Service Trac) for input on the administration and sample selection used. With respect to administration, the 2 month window used for including completed surveys is the standard time period used in the industry as well as the exclusion of residents who have been at the AL less than two weeks. 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.

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	Professional Certification or Recognition Program NCAL Quality Initiative Recognition Prorgram https://www.ahcancal.org/ncal/quality/qualityinitiative/Pages/Recognition-Program.aspx

4a1.1 **For each CURRENT use, checked above (update for maintenance of endorsement), provide:**

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

- Level of measurement and setting

Seventeen national satisfaction vendors have agreed to add the CoreQ to their questionnaires and calculate the measure. The following Customer Satisfaction Vendor are using CoreQ:

- Align
- A Place for Mom
- Bivarus
- Brighton Consulting Group
- Cortex Health Inc.
- The Doug Williams Group, Inc.
- Healthcare Academy (ReadyQ)
- Holleran
- Lighthouse Care Updates
- inQ Experience Surveys
- Market Research Answers (CareSat)
- NRC Health
- Nexus Health Resources
- Pinnacle
- Providigm/abaqis
- Sensight Surveys
- Service Trac

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 AL business in the United States.

A user's manual has been developed and is available on the CoreQ website (coreq.org) for all satisfaction survey vendors to use.

AHCA/NCAL has incorporated the CoreQ into their national Quality Initiative goals. AHCA/NCAL has worked with the above mentioned vendors to automatically upload data into LTC Trend Trackersm, a member benefit for tracking data, to allow members to see their CoreQ data. This has resulted in growing number of members and vendors collecting the data.

4a1.2. 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: AL Family Satisfaction 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.

4a1.3. 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.*)

NCAL is preparing to launch the third iteration of the NCAL Quality Initiative. One of the four goals will continue to be to improve customer satisfaction as measured by the CoreQ. AHCA/NCAL developed an upload and reporting feature within its member data profiling tool, LTC Trend Trackersm, which allows SNFs and ALs 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 Trackersm includes an API for vendors performing the survey on behalf of ALs to upload data, so that the aggregate CoreQ results will be available for providers. Given that LTC Trend Trackersm is the leading method for AHCA/NCAL AL members to profile their quality and other data, the incorporation of CoreQ into LTC Trend Trackersm means it will immediately become the de facto standard for customer satisfaction surveys for the AL industry. AHCA/NCAL continues to work with customer satisfaction vendors to promote CoreQ and receives requests for vendors to be added to the list of those incorporating CoreQ.

The CoreQ team is working with states who require satisfaction measurement to incorporate the CoreQ into their process. AHCA/NCAL has a presence in almost every state, and these state affiliates will be promoting the use of the CoreQ in those states that are collecting or considering collecting satisfaction. AHCA/NCAL has reached out to other provider groups to promote CoreQ to their members.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Not applicable.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
Not applicable.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Not applicable.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Not applicable.

4a2.2.3. **Summarize the feedback obtained from other users**
Not applicable.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
Not applicable.

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.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)**
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 in use for performance improvement. Higher CoreQ Family Satisfaction measure ratings indicate higher quality of health care. Patient/Family preference is key in health care and important for assisted living which was started as a person-centered care model. Focusing on measuring family satisfaction with care is important for providers to include in their quality improvement efforts.

4b2. 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).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
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: AL Family Satisfaction 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 AL 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.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**

Not applicable.

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)**
2614 : CoreQ: Short Stay Discharge Measure
2615 : CoreQ: Long-Stay Resident Measure
2616 : CoreQ: Long-Stay Family Measure
3420 : CoreQ: AL Resident Satisfaction Measure

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
We have developed three skilled nursing facility (SNF) based CoreQ measures (CoreQ: Long-Stay Family Satisfaction Measure, CoreQ: Long-Stay Resident Satisfaction Measure, and CoreQ: Short-Stay Discharge Measure), which are currently used in nursing facilities. These three measures are NQF endorsed as of 2016.

5a. Harmonization of Related Measures
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 harmonized to the extent possible?
Yes

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.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Health Care Association/National Center for Assisted Living
Co.2 **Point of Contact:** Lindsay, Schwartz, lschwartz@ncal.org, 202-898-2848-
Co.3 **Measure Developer if different from Measure Steward:** American Health Care Association/National Center for Assisted Living
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.
AHCA/NCAL Staff including:
David Gifford, MD
Lindsay B. Schwartz, Ph.D.
Nicholas Castle, Ph.D. University of Pittsburgh
Mary Tess Crotty, Genesis
Matt O'Connor, Ph.D., HCR Manor Care (employer at time of development of CoreQ)
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

Dr. Nicholas Castle collected and analyzed data and work with workgroup and team on measure development. The workgroup gave input, reviewing our suggested administration, required response rate, the manual, and exclusions. Dr. O'Connor worked on analyses. Both Dr. O'Connor and Mary Tess provided feedback to AHCA/NCAL staff and Dr. Castle on the manual and helped along the way reviewing analyses and the development process.

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:**