NATIONAL QUALITY FORUM

1030 15TH STREET, NW, SUITE 800
WASHINGTON, DC 20005
202 783-1300 MAIN 783-3434 FAX

TO: Composite Expert Panel

FR: Helen Burstin, Karen Pace, Karen Johnson, Elisa Munthali

SU: Briefing materials for November 15, 2012 conference call

DA: November 12, 2012

Thank you for your outstanding discussions at the in-person meeting. The purpose of this call is to:

- review and modify draft principles and evaluation criteria for composite performance measures;
- identify any outstanding issues; and
- make suggestions for format and content of the draft report.

The draft principles and recommendations included in this memo represent our first attempt to synthesize the discussions from the in-person meeting. Because these principles and recommendations will form the backbone of our report for this project, we look forward to your review and suggestions.

Expert Panel Action

- Review this briefing memo and background materials posted on SharePoint at: http://share.qualityforum.org/Projects/Composite%20Measures%20Guidance%20Evaluation/SitePages/Home.aspx

Conference Call Information

Date/Time: Thursday, November 15, 3:00-5:00 pm ET
Title: Composite Measure Evaluation Guidance Expert Panel – Conference Call
Telephone dial-in #: (888) 799-5160
Confirmation code: 33349337
Weblink: http://nqf.commpartners.com/se/Rd/Mt.aspx?547763
You will be prompted to enter your name, location (optional), and e-mail address. Then click on "Click here to enter presentation."
For technical support, please e-mail nqf@commpartners.com.

CONTENTS

CLARIFICATIONS

In reviewing the transcript of the meeting, we noted a few areas where terminology was confusing. To ensure common understanding, NQF staff will include a glossary in the draft report.

Evidence

The term "evidence" was used to refer to NQF's evidence criterion, which is focused on the clinical evidence of the measure focus, as well as evidence for other criteria such as reliability and validity. NQF's evidence criterion refers to the clinical evidence for the measure focus (e.g., the evidence that maintaining blood pressure below 140/90 is associated with lower mortality or morbidity). NQF also requires evidence related to measurement science (i.e., reliability and validity of the performance measure as constructed).

Instrument vs. Performance Measure

Instruments such as the CAHPS or PHQ-9 are used to collect data at the patient level. Some instruments also may be referred to as composites. Data from such instruments may be used in performance measures that aggregate data for all patients of a healthcare provider (see PRO report).

NQF Current Evaluation Criteria

Since the initial composite report was released in early 2010, NQF has updated its criteria for evidence, measure testing, and usability. The latest version of the evaluation criteria are included in Table 1 and Appendix A.

PRINCIPLES

The following key principles were identified from the TEP discussions during the in-person meeting; these principles guided the TEP's recommendations regarding the evaluation criteria.

- The term "composite measure" may be applied to many types of measures, including individual-level instruments and performance measures. NQF only endorses performance measures.
- Approaches to composite measure development and construction are described using a variety of terms and can vary by discipline. Nonetheless, the construct and evaluation of composite measures

should be based on sound measurement science principles, and discipline-specific jargon (such as "psychometric" and "clinimetric") should be avoided.

- The quality construct and purpose of the composite performance measure are essential for determining what components are included in a composite performance measure and what analyses should be used to demonstrate reliability and validity.
- Composite performance measures should provide an added value over that of individual measures alone.
- A desire to create one score from multiple performance measures is not a sufficient quality construct or purpose for creating a composite performance measure.
- NQF-endorsement of the individual component measures should not be mandatory; however, NQF endorsement of the component measures could satisfy some requirements for the component measures included in a composite.
- The individual components that are included in a composite performance measure should be justified based on the clinical evidence (i.e., what is being measured is based on clinical evidence of a link to desired outcomes).
- The individual components in a composite performance measure generally should demonstrate a gap in performance; however, there may be conceptual or analytical justification for including components that do not have a gap in performance.
- The individual components in a composite performance measure may or may not be correlated, depending on the quality construct.
- The reliability of the composite performance measure is of greater interest than the reliability of the individual components. The individual components do not have to be independently reliable.
- The validity of the composite performance measure is of greater interest than the validity of the individual components. Even if the components are valid measures, the construction of the composite may not be a valid representation of the quality construct.
- When evaluating composite performance measures, Steering Committees should discuss both the quality construct itself as well the empirical evidence for the composite (i.e., supporting the method of construction and methods of analysis).
- Components should be "necessary"—either empirically (i.e., they contribute to the reliability) or conceptually.
- Composite performance measures are complex and attention should be given to parsimony and simplicity.

DEFINITION

A composite [performance] measure is a combination of two or more individual [performance] measures in a single measure that results in a single score.

EVALUATION CRITERIA

In the following table, the left column includes the latest 2012 NQF measure evaluation criteria, which incorporates the recent changes to Usability and Use. The notes are hyperlinked to the criteria plus notes in Appendix A. The middle column includes the additional criteria for composites from the 2008/2009 guidance document; these have been redlined to show draft changes per the TEP discussions. The right column includes rationale and outstanding questions.

Table 1. Evaluation Criteria

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
1.Evidence, Performance Gap, and 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 aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to*** *meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.*	~~The individual measures included in the composite or subcomposite measures must be either: NQF endorsed *OR* assessed to have met the individual measure evaluation criteria as the first step in evaluating the composite measure. (This does not apply to subscales of a scale/ instrument that cannot be used independently of the total scale.)~~	This statement is no longer needed with the following guidance.
	~~If the component measures are determined to meet the importance criteria 1a, 1b, and 1c, then the composite would meet 1a, 1b, and 1c. A component measure might not be important enough in its own right as an individual measure, but it could be determined to be an important component of a composite.~~	This statement is no longer relevant with the following guidance.
1a. Evidence to Support the Measure Focus The measure focus and components of a composite ~~is~~are evidence-based, demonstrated as follows: • Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. • Intermediate clinical outcome: 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.	**Guidance for Composite:** The evidence criterion (1a) must be met for each component (unless component is NQF-endorsed under the new evidence requirements).	

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions

• 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.
• Experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
• Efficiency: [6] evidence not required for the resource use component.
AND
1b. Performance Gap
Demonstration of quality problems and opportunity for improvement, i.e., data [7] demonstrating
  • considerable variation, or overall less-than-optimal performance, in the quality of care across providers; **and/or**
  • disparities in care across population groups.
AND
1c. High Priority
The performance measure addresses:
• a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
OR
• a demonstrated high-priority aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).
	The performance gap criterion (1b) must be met for each component, and if possible, for the composite performance measure as a whole. If a component measure has little opportunity for improvement, justification for why it should be included in the composite is required (e.g., increase reliability of the composite, clinical evidence).
	
	The priority criterion (1c) applies to the composite performance measure as a whole.
	
	Composite. 1d. ~~The purpose/objective of the composite measure and the construct for quality are clearly described.~~ The following must be clearly articulated for the composite performance measure:
	• The quality construct
	• The purpose, including how the composite measure provides a distinctive or additive value and	

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
	better achieves the purpose than do the components individually. • How the methods for development and the components that are used to construct the composite are consistent with and representative of the stated quality construct and purpose.	Is the last bullet necessary if we have identified the correct criteria under scientific acceptability? (How would you demonstrate that the components are representative of the quality construct? – evidence for each component or some type of analysis under testing? How would you demonstrate it is consistent with purpose?)
	~~Composite. 1c. The component items/ measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure. Whether the composite measure development begins with a conceptual construct or a set of measures, the measures included must be conceptually coherent and consistent with the purpose.~~	This is included in 1d.
2. Reliability and Validity—Scientific Acceptability of Measure Properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to*** *meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.* **2a. Reliability** **2a1.** The measure is well defined and precisely specified [8] so it can be implemented consistently	**Guidance for Composite Performance Measures 2a-2c** 2a1. ~~The composite measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and~~	

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
within and across organizations and allows for comparability. EHR measure specifications are based on the quality data model (QDM). [9]	~~allow for comparability.~~ Composite specifications include methods for standardizing scales across component scores, scoring rules (i.e., how the component scores are combined or aggregated), weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite), handling of missing data, and required sample sizes.	
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.	~~2b. Reliability testing of the composite measure demonstrates that the results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.~~ 2a2. For composite performance measures, reliability must be demonstrated for the measure score. If the components will be disaggregated, then reliability for the component measures must be demonstrated (unless they are NQF-endorsed).	Are there any circumstances where reliability of data elements or reliability of the individual performance measure scores is acceptable?
2b. Validity **2b1.** The measure specifications [8] are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.	~~2c. Validity testing demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed.~~	Are there any circumstances where validity of data elements or validity of the individual performance measure scores is acceptable?
2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.	2b2. For composite performance measures, validity must be demonstrated for the measure score. If the components will be disaggregated, then validity for the component measures must be demonstrated (unless they are NQF-endorsed).	Are there any circumstances where face validity Is acceptable for the composite performance measure?

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12] **AND** If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]	2b3. Exclusions apply primarily to the component measures. It would not need to be addressed if validity of the composite performance measure was demonstrated.	
2b4. For outcome measures and other measures when indicated (e.g., resource use): • an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration **OR** • rationale/data support no risk adjustment/ stratification.	2b4. This would be required for outcome component measures (unless they are NQF-endorsed).	
2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance; **OR** there is evidence of overall less-than-optimal performance.	2b5. Applies to composite performance measures.	
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.	2b6. Applies to component measures.	
2c. Disparities If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity,	2c. Applies to composite performance measures.	

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
socioeconomic status, gender); **OR** rationale/data justifies why stratification is not necessary or not feasible.	**2d. For composite performance measures, analyses support that the specified scoring/weighting and included component measures are consistent with and representative of the stated quality construct and purpose with adequate attention to parsimony and simplicity.**	Can 2i-2l be incorporated as one criterion for composite measures (similar to 2b4 for outcome measures)? For example, see 2d.
	Composite. 2i. Component ~~item/~~measure analysis ~~(e.g., various correlation analyses such as internal consistency reliability),~~ demonstrates that the included component ~~items/~~measures fit the ~~conceptual~~ quality construct~~; OR justification and results for alternative analyses are provided~~.	What would these analyses be?
	~~Composite. 2j. Component item/measure analysis demonstrates that the included components contribute to the variation in the overall composite score; OR if not, justification for inclusion is provided.~~ For component measures: Either empirical or conceptual justification must be provided to demonstrate that adequate attention has been paid to parsimony.	What would this be?
	Composite. 2k. The scoring/ aggregation and weighting rules are consistent with the ~~conceptual~~ quality construct, with a preference for simplicity and ease of presentation, and must be justifiable, preferably though empirical analysis. ~~(Simple, equal weighting is often preferred unless differential weighting is justified.~~	

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
	Composite. 2l. Analysis of missing data for each component ~~scores supports the specifications for scoring/ aggregation and handling of missing component scores~~ includes a quantification of missing data and how the specified treatment of missing data minimizes bias.	Can missing data be incorporated into criterion regarding exclusions?
3. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement. **3a.** For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order). **3b.** 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. **3c.** Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, [17] 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).	3a, 3b, 3c. Apply to composite performance measures. ~~a. For clinical composite measures, overall the required data elements are routinely generated concurrent with and as a byproduct of care processes during care delivery.~~ ~~b. The required data elements for the composite overall are available in electronic sources.~~ ~~c. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, etc.) for obtaining all component measures can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational.~~	No need to repeat criteria when apply to composites. NOTE: How will patient-reported data for PRO-PMs be addressed?
4. Usability and Use Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement [18] to achieve the goal of high-quality, efficient healthcare for individuals or populations. **4a. Accountability and Transparency** [19]	~~3a. Demonstration that information produced by the composite measure~~	This was the old usability language.

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
Performance results are used in at least one accountability application [1] within three years after initial endorsement and are publicly reported [19] within six years after initial endorsement (or the data on performance results are available). [20] If not in use at the time of initial endorsement, then a credible plan [21] for implementation within the specified timeframes is provided. **AND** **4b. Improvement** [22] Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. [22] 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. **AND** **4c.** 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).	~~is meaningful, understandable, and useful to the intended audience(s) for both public reporting (e.g., focus group, cognitive testing) and informing quality improvement (e.g., quality improvement initiatives).~~ 4a. Applies to composite performance measures. ~~Composite. 3e. Demonstration (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective.~~ 4b. Applies to composite performance measures, except it should explicitly link back to the quality construct and purpose. ~~d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.~~ 4c. Applies to composite performance measures. If there is evidence of unintended negative consequences for one of the components, the developer should explain how that is handled or justify why that component should remain in the composite. ~~Composite. 3d. Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.~~ 4d. To facilitate transparency and understanding, data should be collected and components composited in such a way as to permit disaggregation into component scores.	No longer needed if addressed under 4b improvement This was old language. Is this a criterion? How do we evaluate? Does it fit with Usability and Use?
5. Comparison to Related or Competing Measures	~~3b. The component measure specifications are harmonized.~~	Composite performance

2012 Endorsement Criteria for All Measures	Composite Criteria- Draft Changes	Notes/Questions
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. **5a.** The measure specifications are harmonized [23] with related measures; **OR** the differences in specifications are justified. **5b.** The measure is superior to competing measures (e.g., is a more valid or efficient way to measure); **OR** multiple measures are justified.	~~3c. Review of existing endorsed measures and measure sets demonstrates that the composite measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of healthcare, is a more valid or efficient way to measure).~~ 5a and 5b. Applies to composite performance measures and the component measures.	measures are subject to basic criteria regarding related and competing measures.

ADDITIONAL RECOMMENDATIONS

- Steering Committees should include at least one member who is knowledgeable about composite measures and/or composite measures should undergo a methodological technical expert consultation
- Provide examples of types of analyses for different types of composite performance measures

ADDITIONAL QUESTIONS FOR TEP INPUT

- If the definition doesn't change, how do developers, staff, committees identify what measures are subject to additional criteria for composite performance measures?
- Does the principle of increased reliability with increased number of items hold for all-or-none measures when multiple components are reduced to one data point?
- Do analyses such as factor analysis and internal consistency reliability need any modification when the unit of analysis is providers (vs. people) and the data are performance measure scores (vs. item responses)?

APPENDICES

Appendix A—Measure Evaluation Criteria

NATIONAL QUALITY FORUM

DRAFT 11/10/2012 after CSAC

Measure Evaluation Criteria

Effective November 2012

Conditions for Consideration
Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**

A. The measure is in the public domain or a measure steward agreement is signed.

B. The measure owner/steward verifies there is an identified responsible entity and a process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every three years.

C. The intended use of the measure includes both accountability applications [1] (including public reporting) and performance improvement to achieve high-quality, efficient healthcare.

D. The measure is fully specified and tested for reliability and validity. [2]

E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.

F. The requested measure submission information is complete and responsive to the questions so that all the information needed to evaluate all criteria is provided.

Notes
1. Accountability applications are the use of performance results about identifiable, accountable entities to make judgments and decisions as a consequence of performance, such as reward, recognition, punishment, payment, or selection (e.g., public reporting, accreditation, licensure, professional certification, health information technology incentives, performance-based payment, network inclusion/exclusion). **Selection** is the use of performance results to make or affirm choices regarding providers of healthcare or health plans.

2. A measure that has not been tested for reliability and validity is only potentially eligible for time-limited endorsement if all of the following conditions are met: 1) the measure topic is not addressed by an endorsed measure; 2) it is relevant to a critical timeline (e.g., legislative mandate) for implementing

endorsed measures; 3) the measure is not complex (requiring risk adjustment or a composite); and 4) the measure steward verifies that testing will be completed within 12 months of endorsement.

Criteria for Evaluation
If all conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Usability*, and *Feasibility*. Not all acceptable measures will be equally strong among each set of criteria. The assessment of each criterion is a matter of degree. However, if a measure is not judged to have met minimum requirements for *Importance to Measure and Report* or *Scientific Acceptability of Measure Properties*, it cannot be recommended for endorsement and will not be evaluated against the remaining criteria.

1.Evidence, Performance Gap, and 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 aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to*** *meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.*

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care.
- 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.
- Experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency: [6] evidence not required for the resource use component.

AND

1b. Performance Gap
Demonstration of quality problems and opportunity for improvement, i.e., data [7] demonstrating

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

AND

1c. High Priority
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;

OR

- a demonstrated high-priority aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

Notes
3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.
6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).
7. Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

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

2a. Reliability
2a1. The measure is well defined and precisely specified [8] so it can be implemented consistently within and across organizations and allow for comparability. EHR measure specifications are based on the quality data model (QDM). [9]

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.

2b. Validity
2b1. The measure specifications [8] are consistent with the evidence presented to support the focus of measurement under criterion 1c. The measure is specified to capture the most inclusive target

population indicated by the evidence, and exclusions are supported by the evidence.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

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

AND

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

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

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

2c. Disparities

If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender);

OR

rationale/data justifies why stratification is not necessary or not feasible.

Notes

8. Measure specifications include the target population (denominator) to whom the measure applies,

identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation.
9. EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

3. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3b. 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.

3c. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, [17] 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).

Note
17. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

4. Usability and Use
Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement [18] to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency [19]
Performance results are used in at least one accountability application [1] within three years after initial endorsement and are publicly reported [19] within six years after initial endorsement (or the data on performance results are available). [20] If not in use at the time of initial endorsement, then a credible plan [21] for implementation within the specified timeframes is provided.

AND

4b. Improvement [22]
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. [22] 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.

AND

4c. 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).

Notes
18. An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.
19. Transparency is the extent to which performance results about identifiable, accountable entities are *disclosed and available* outside of the organizations or practices whose performance is measured. Maximal transparency is achieved with **public reporting** defined as making comparative performance results about identifiable, accountable entities freely available (or at nominal cost) to the public at large (generally on a public website). *At a minimum, the data on performance results about identifiable,*

accountable entities are available to the public (e.g., unformatted database). The capability to verify the performance results adds substantially to transparency.
20. This guidance is not intended to be construed as favoring measures developed by organizations that are able to implement their own measures (such as government agencies or accrediting organizations) over equally strong measures developed by organizations that may not be able to do so (such as researchers, consultants, or academics). Accordingly, measure developers may request a longer timeframe with appropriate explanation and justification.
21. 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.
22. Demonstrated progress toward achieving the goal of high-quality, efficient healthcare includes evidence of improved performance and/or increased numbers of individuals receiving high-quality healthcare. Exceptions may be considered with appropriate explanation and justification.

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.

5a. The measure specifications are harmonized [23] with related measures;

OR

the differences in specifications are justified.

5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

multiple measures are justified.

Note
23. Measure harmonization refers to the standardization of specifications for related measures with the same measure focus (e.g., *influenza immunization* of patients in hospitals or nursing homes); related measures with the same target population (e.g., eye exam and HbA1c for *patients with diabetes*); or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

Appendix B—Glossary

Term	Definition	Source
All-or-None Scoring *Also known as:* • *Appropriateness model* • *Conjunctive scoring*	A percentage is determined by applying an all-or-none rule at the patient level. The denominator is the number of patients eligible to receive at least one of the identified elements of care, and the numerator is the number of patients who actually received all of the care for which the specific patient was eligible. No partial credit is given.	NQF Composite Guidance Report, 2007
Bundle	A series of interventions related to a specific condition that, when implemented together, will achieve significantly better outcomes than when implemented individually. This term was developed by faculty at the Institute for Healthcare Improvement. See www.ihi.org/IHI/Topics/CriticalCare/IntensiveCare/ImprovementStories/BundleUpforSafety.htm.	NQF Composite Guidance Report, 2007
Clinimetric approach *Will be updated based on this project*	Approach to developing a scale that relies on the required relationships between the observed items and the attribute for which an index is being defined. The most important attributes to be included in the index are not expected to be homogeneous because they indicate different aspects of a complex clinical phenomenon.	NQF Composite Guidance Report, 2007F
Component	A constituent part or element of a composite measure.	NQF Composite Guidance Report, 2007
Composite measure	A combination of two or more individual measures into a single measure that results in a single score.	NQF Composite Guidance Report, 2007
Construct	An abstract phenomenon that is measured indirectly through less abstract indicators.	NQF Composite Guidance Report, 2007
Domain	A dimension or aspect of a construct.	NQF Composite Guidance Report, 2007
Indicator	Sometimes used interchangeably with measure, but may indicate a more descriptive level than the term “measure,” which indicates the operational definition.	NQF Composite Guidance Report, 2007
Indicator Average	For each indicator, the percentage of times the indicator was met is computed. The scores are averaged across all indicators. This score represents the mean rate at which each audited aspect of care was met.	Reeves, 2007
Item	A single question on a measurement scale or instrument	NQF Composite Guidance Report, 2007
Latent variable	An unobserved trait or characteristic	NQF Composite

Term	Definition	Source
		Guidance Report, 2007
Measure	Numeric quantification of some concept. A quality measure is a numeric quantification of healthcare quality.	NQF Composite Guidance Report, 2007
Opportunity scoring	Scoring used with process measures, determined from the sum of all numerators (achieved the desired process) divided by the sum of all denominators (i.e., number of eligible patients or opportunities, which could vary by measure). If the opportunity score is based on "care events" (patient/provider interactions), the opportunity score is the percentage of all care events that were met. For example, if patient A meets 1 of 1 opportunity and patient B meets 3 of 4 opportunities, then the care event opportunity score =80% [i.e., (1+3)/(1+4)]. If the opportunity score is based on patients, the opportunity score is some function (typically the average) of the number of care events that were met for each patient. Using the above example, the patient-based opportunity score =88% [i.e., 100% met for patient A, 75% met for patient B→average over the 2 patients= 100+75 / 2. (Has also been called "patient average".)	NQF, Composite Guidance Report, 2007, Aligning Forces, 2010, Reeves, 2007
Paired measures	Individual measures that should be measured concurrently in the same population; however, the results are not combined into a single score.	NQF Composite Guidance Report, 2007
Percentage Standard	This is a less stringent version of the All-or-None method, where the criterion for success is that some percentage (e.g., 70%) or more of the triggered indicators be met.	Reeves, 2007
Psychometric approach *Will be updated based on this project*	Approach to developing a scale that relies on the relationships between the items that have been measured where the multiple component items are all measuring more or less the same single attribute.	NQF Composite Guidance Report, 2007
Scale	A measure of an attribute composed of a set of related items. A score on the scale represents a point along a continuum representing more or less of the attribute.	NQF Composite Guidance Report, 2007
Subscale	A measure of a dimension of a scale composed of a subset of the items in a scale.	NQF Composite Guidance Report, 2007
Variable	A characteristic or attribute that varies within and among people or the subjects of study.	NQF Composite Guidance Report, 2007