NATIONAL QUALITY FORUM

Memo

TO: NQF Members and Public

FR: NQF Staff

RE: Review of *Composite Performance Measure Evaluation Guidance*

DA: November 29, 2012

Background

Composite performance measures, which combine information on multiple individual measures into one single measure, are of increasing interest in healthcare performance measurement and public accountability applications. Such measures are complex and require a strong conceptual and methodological foundation with different considerations for testing and analysis.

NQF previously developed guidance to assist steering committees with their evaluation of composite performance measures as part of the NQF endorsement process. Since that time, NQF has gained experience with composite performance measures and has identified some challenges and issues with implementing the prior guidance. Additionally, NQF has updated the standard measure evaluation criteria and guidance for evidence, measure testing, and usability; thus, there also is a need to align the evaluation criteria for composite measures with the updated guidance.

The purpose of the Composite Performance Measure Evaluation Guidance Project was to review and update NQF's criteria and guidance on evaluating composite performance measures for potential NQF endorsement. NQF convened a 12-member Technical Expert Panel (TEP) to achieve the following goals of the project:

- review the existing guidance for evaluating composite performance measures;
- identify any unique considerations for evaluating composite performance within the context of NQF's updated endorsement criteria; and
- modify existing criteria and guidance and/or provide additional recommendations for evaluating composite performance measures.

Review and Comment

The TEP's recommendations are included in the draft document *Composite Performance Measure Evaluation Guidance*. The draft report is posted on the NQF web site for purposes of review and comment only and is not intended to be used for voting purposes.

The recommendations include modifications to the current NQF criteria and guidance for evaluating composite performance measures. Of particular note is the elimination of the requirement that component performance measures must be NQF-endorsed or meet the criteria for endorsement. Instead, the recommended guidance indicates which subcriteria should be met for the component measures (e.g., evidence and performance gap) and which should be met for the composite measure (e.g., reliability and validity). There are two

subcriteria specific to composite performance measures that are incorporated into the standard criteria. Under Importance to Measure and Report, subcriterion 1d requires a clear and logical description of the quality construct, rationale, and how the composite measure construction is consistent with them. Under Scientific Acceptability of Measure Properties, subcriterion 2d requires analyses to demonstrate that the component measures fit the quality construct, the aggregation and weighting rules fit the quality construct and rationale, and extent and handling of missing data.

The TEP recommended that NQF develop examples of types of analyses that are appropriate for various approaches to composite measure construction. NQF staff will continue to work with the TEP to develop Appendix B as a resource for steering committees, staff, and measure developers. Suggestions are welcome.

You may post your comments and view the comments of others on the NQF website using the online submission process.

All comments must be submitted no later than 6:00 PM ET, December 28, 2012.

Thank you for your interest in the NQF's work. We look forward to your review and comments.

Composite Performance Measure Evaluation Guidance

DRAFT REPORT FOR REVIEW

November 29, 2012

NATIONAL QUALITY FORUM

Contents

Composite Performance Measure Evaluation Guidance

DRAFT REPORT

Introduction

Healthcare is a complex and multidimensional activity. While individual performance measures provide much important information, there is also value in summarizing performance on multiple dimensions. Composite performance measures, which combine information on multiple individual measures into one single measure, are of increasing interest in healthcare performance measurement and public accountability applications. According to the Institute of Medicine,[1] such measures can enhance the performance measurement enterprise and provide a potentially deeper view of the reliability of the care system. Further, composite performance measures may be useful for multiple stakeholders, including consumers, purchasers, and policy makers.

Composite performance measures are complex and require a strong conceptual and methodological foundation with different considerations for testing and analysis. As with individual performance measures, the methods used to construct composite performance affects the reliability, validity, and usefulness of the composite measure.

Several composite measures are included in NQF's portfolio of endorsed measures, and NQF previously developed guidance[2] to assist steering committees with their assessment of these measures as part of the NQF evaluation process. Since that time, however, NQF has updated the standard measure evaluation criteria and guidance for evidence, measure testing, and usability; thus, there is a need to align the evaluation criteria for composite measures with the updated guidance.

Purpose

The purpose of the Composite Performance Measure Evaluation Guidance Project was to review and update NQF's criteria and guidance on evaluating composite performance measures for potential NQF endorsement. Specifically, the goals of the project were to:

- review the existing guidance for evaluating composite performance measures;
- identify any unique considerations for evaluating composite performance within the context of NQF's updated endorsement criteria;
- modify existing criteria and guidance and/or provide additional recommendations for evaluating composite performance measures.

To achieve these goals, NQF convened a 12-member Technical Expert Panel (TEP), which was comprised primarily of methodologists and other experts in the development of composite performance measures. In addition to participating in several conference calls, the TEP also gathered for a one-day in-person meeting in Washington, DC on November 2, 2012.

Background

Prior Guidance on Evaluating Composite Measures

In 2008-2009, NQF initiated a project to identify a framework for evaluating composite performance measures. That developmental work included defining composite performance measures, articulating principles underlying the evaluation of composite performance measures, and developing an initial set of specific criteria (to be used in addition to NQF's standard evaluation criteria) with which to evaluate composite performance measures for potential NQF endorsement.

The principles articulated for evaluating composite performance measures reflected the need for a conceptual construct of quality underlying the composite measure and justification of the methods used to construct and test the measure for reliability and validity. The criteria emphasized the need for transparency around the methodology used for composite measure construction and required that both the components of the composite and the composite measure as a whole meet NQF's measure evaluation criteria. This work served as the basis for the current project.

NQF Experience with Composite Performance Measures

Since 2007, 28 measures submitted to NQF for potential endorsement have been flagged as composite measures. Of these, 22 are currently endorsed. The majority of the endorsed composite measures (n=11) are derived from surveys targeted towards patients or consumers (e.g., the Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys). The remainder of the endorsed composite performance measures are comprised of all-or-none measures (n=3) and composites constructed using various methods of aggregation and weighting methodologies (n=8). As with NQF-endorsed individual measures, these composite measures are considered suitable both for performance improvement and accountability applications.

However, the evaluation of composite measures for potential endorsement has not been without difficulty. The most common issues have revolved around the identification of composite measures, ambiguity in the guidance when a component measure is not NQF-endorsed, and incomplete submissions.

Identifying Measures as Composites

Not all composite measures that have met—or potentially have met—the current NQF definition of composite measures have been flagged by the measure developers as composite measures. These include all-or-none composites in which the components are assessed at the patient level (i.e., whether each patient received each required process then aggregated for the healthcare entity); simpler all-or-none measures that require multiple conditions (e.g., assess vaccination status and administer flu vaccine); and any-or-all measures that assess whether a patient has exhibited any or all of a list of complications. For such measures, it is unclear whether the additional analyses indicated for composite measures (e.g., analysis of components to demonstrate alignment with the conceptual construct and contribution to the variation in the overall composite score) are applicable, and often these additional analyses have not been submitted by developers of these measures.

Evaluation of Component Measures

The current guidance indicates that the component measures that make up a composite measure should be NQF-endorsed or evaluated as meeting the individual measure evaluation criteria as the first step in evaluating the composite measure. However, the guidance goes on to state that while a component measure might not be important enough in its own right as an individual measure, it could be determined to be an important component of a composite. Some developers have interpreted this guidance to mean that components do not need to meet the Importance to Measure and Report criteria around evidence, impact, and performance gap. But this interpretation regarding evidence and performance gap calls into question the basis for including the component measure. Another issue related to the evidence criterion is whether measures that are distal to desired outcomes could be included in composite measures. For example, a performance measure of merely obtaining a lab test is not considered to meet the criteria because it is so distal to the desired outcome and is often based on expert opinion; however, this type of component has been suggested for inclusion in a composite measure.

It also is not clear whether balancing measures that would not meet the importance criteria should be included in a composite performance measure. A balancing measure is not the main focus of interest but is used to identify or monitor potential adverse consequences of measurement. For example, a performance measure about treating substance use that requires the identification of patients with substance use problems will not be accurate if most patients are not even screened; therefore, a screening measure might be considered a balancing measure.

Finally, it has been difficult to apply criteria for related and competing measures to composite measures. The challenges with measure harmonization are amplified with composite measures because, typically, more measures (involving multiple developers) will be involved in harmonization discussions. While using previously-endorsed measures as components in a composite measure should ameliorate most difficulties around harmonization, often the components in submitted composite measures have not been previously endorsed. In such cases, these components either competed directly with other endorsed measures or were not harmonized with endorsed measures.

Incomplete Submissions Related to Requirements for Composite Measures

As discussed earlier, if measures are not flagged as composite measures, then the additional information needed to evaluate them as composite measures may not be submitted by measure developers. However, non-responsiveness to composite-specific items also has been a problem. For example, the current criteria state that the purpose/objective of the composite measure and the construct for quality must be clearly described, yet often little beyond a list of the component measures is provided.

Current criteria require testing for reliability and validity of the composite measure (even if the individual measures have demonstrated reliability and validity), as well as additional analyses to justify the inclusion of component measures and the specified aggregation and weighting rules. Reliability and validity testing of the composite measure may not be conducted. Some of the composite questions refer to correlational analyses, which may not be appropriate for all composite measures. While the current guidance recognizes this and indicates that the developer could submit other analyses with rationale, these alternative analyses have not always been submitted (or if submitted, the rationale may not have been included or may not have been sufficiently explanatory). Analysis of the contribution of individual components to the composite score often has not been submitted. Without this information, steering

committees may be left with little more than face validity as a basis for recommending a composite performance measure.

Definition of Composite Performance Measure

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

The TEP reviewed and retained the definition provided in the initial composite report and added explicit clarification that it refers to composite *performance* measures. Note that the term "composite measure" also refers to individual-level measures (i.e., instruments and scales used to obtain data from individuals, such as the CAHPS or PHQ-9). Data from such instruments may be used in performance measures that aggregate data for all patients served by a healthcare entity, but that does not in itself make it a composite performance measure. Patient-reported outcomes (PRO) and performance measurement has been the subject of a recent NQF project (see PRO report). Throughout this report, the terms composite measure or component measure refer to a performance measure unless otherwise indicated.

Types of Composite Measures

Composites often are classified according to the empirical and conceptual relationship among the component measures and between the components and the composite, the rules for combining the individual components (e.g., all-or-none, opportunity, weighted average), or the type of individual measures included in the composite (e.g., process, outcome). Regardless of the various approaches, a coherent quality construct and rationale should guide the composite development and testing and analysis. The glossary in Appendix A contains definitions for various approaches to combining the component measures. Appendix B provides a description of various conceptual models for the relationship among the component measures and composite score and identifies relevant analyses. The TEP suggested that over time, NQF add specific examples of composite performance measures that use these conceptual models.

The TEP decided that a classification of types of composite performance measures would not be particularly useful and could lead to unnecessary attention to the approach used to construct the composite. They agreed that the primary concern for NQF endorsement is whether the resulting composite performance measure is based on sound measurement science, produces a reliable signal, and is a valid reflection of quality.

Key Steps in Developing a Composite Performance Measure

A variety of methods can be used to construct composite performance measures; however, they all involve the following key steps: [3-9]

- Describing the quality construct to be measured and rationale for the composite;
- Selecting the component measures to be combined in the composite measure;
- Ensuring that the methods used to aggregate and weight the components supports the goal that is articulated for the measure;
- Combining the component measure scores, using the specified method; and

- Testing the composite measure to determine if it is a reliable and valid indicator of quality healthcare.

Guiding Principles

The following key principles were identified by the TEP and guided their recommendations regarding the evaluation criteria.

Terminology

As noted above, the TEP opted for a broad, generic definition of composite performance measure.

- The term "composite measure" may be applied to many types of measures, including individual-level instruments as well as aggregate-level 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 construction and evaluation of composite performance measures should be based on sound measurement science principles. Although often used in the published literature on composite measures, the TEP wanted to minimize the use of discipline-specific language and categorizations (e.g., "psychometric" and "clinimetric") in the evaluation criteria and guidance.

Component Performance Measures

The prior composite evaluation criteria required that each component performance measure be NQF-endorsed or meet all criteria for NQF endorsement. At times that has been difficult to implement, particularly for reliability. The TEP noted that individual measures may not be reliable independently because of rare events or small case volume, but could be used successfully within a composite because the composite combines multiple measures, which can increase reliability of the composite performance measure as a whole. Rather than requiring that each component meet all NQF criteria, the TEP focused on the overall composite and identified those NQF criteria that must be met to justify inclusion of the individual component measures. The TEP agreed, however, that if an individual component measure is NQF-endorsed, then those criteria would not need to be demonstrated again.

- 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 component measures that are included in a composite performance measure should be justified based on the clinical evidence (i.e.., for process measures, what is being measured is based on clinical evidence of a link to desired outcomes; for health outcomes, a rationale that it is influenced by healthcare).
- 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 may not be sufficiently reliable independently, but could contribute to the reliability of the composite performance measure.

Composite Performance Measure

The TEP emphasized the need for a coherent quality construct and rationale to guide construction of the composite as well as to guide evaluation for NQF endorsement. A quality construct is a hypothetical complex concept of quality. Component measures should be selected based on fit with the quality construct, and analyses should justify that fit. All composite performance measures share the potential for simplification when representing one score versus many scores for individual performance measures. However, that feature alone is not sufficient justification for a composite performance measure. Each component should fit the construct and be necessary. The composite performance measure should provide added value over having individual performance measures. Composite measures are complex with aggregation and weighting rules that are not applicable to the individual component measures; therefore, reliability and validity of the composite performance measure score should be demonstrated.

- A coherent quality construct and rationale for the composite performance measure are essential for determining:
    - what components are included in a composite performance measure;
    - how the components are aggregated and weighted;
    - what analyses should be used to support the components and demonstrate reliability and validity; and
    - added value over that of individual measures alone.
- Reliability and validity of the individual components do not automatically ensure reliability and validity of the constructed composite performance measure. Reliability and validity of the constructed composite performance measure should be demonstrated.
- When evaluating composite performance measures, both the quality construct itself, as well the empirical evidence for the composite (i.e., supporting the method of construction and methods of analysis), should be considered.
- Components of a composite performance measure should be "necessary"—either empirically (i.e., they contribute to the reliability) or conceptually. A secondary objective is parsimony, when possible.
- The individual components in a composite performance measure may or may not be correlated, depending on the quality construct.
- Aggregation and weighting rules for constructing composite performance measures should be consistent with the quality construct and rationale for the composite. A secondary objective is simplicity, when possible.
- The standard NQF criteria apply to composite performance measures.

Recommendations for Composite Measure Evaluation

The NQF measure evaluation criteria apply to composite performance measures and their component performance measures. The goal is to incorporate evaluation of composite performance measures into the standard NQF criteria and processes to the extent possible. NQF endorsement is not required for the component measures unless they are intended to be used independently to make judgments about performance. However, the individual component measures should meet specific subcriteria such as for clinical evidence and performance gap, although there may be potential exceptions. The TEP agreed that two additional criteria are needed to evaluate composite performance measures; these are incorporated into the evaluation criteria in Table 1 (see 1d and 2d).

It is important to note the difference between the NQF criteria for evidence and validity. The evidence subcriterion is included under the Importance to Measure and Report criterion and addresses the empirical clinical evidence linking processes to desired health outcomes. In contrast, the validity subcriterion is included under the Scientific Acceptability of Measure Properties criterion and addresses whether the performance measure *as constructed* is an accurate reflection of quality. The clinical evidence provides a justification for measurement and a foundation for validity, but the actual performance measure should be empirically tested to demonstrate validity because how a measure is constructed can affect whether it is an accurate reflection of quality.

Importance to Measure and Report

Evidence

Each component measure must meet the evidence criterion to justify its inclusion in the composite. As with individual performance measures, the evidence requirement ensures that efforts for measurement are devoted to health outcomes or processes of care that will influence desired outcomes. If a component measure is NQF-endorsed (since the updated evidence requirements were implemented), it could be considered as meeting the evidence criterion. If all component measures do not meet the evidence criterion, or do not qualify for the exceptions to the evidence criterion, the composite would not meet the criterion for Importance to Measure and Report unless those components were removed. Evidence is required regardless of approach to constructing a composite measure (i.e., all-or-none scoring or combining scores from individual performance measures).

Performance Gap

Each component measure also should meet the criterion of performance gap to justify its inclusion in the composite. As with individual performance measures, effort for measurement should be directed to aspects of care where there is variation or overall poor performance. However, the TEP acknowledged there may be circumstances when a component measure that does not meet the performance gap criterion could be included in a composite. In such cases, justification for including such a component would be required (e.g., it contributes to the reliability of the overall composite score or is needed for face validity). Ideally, the composite performance measure as a whole also should demonstrate a performance gap.

Quality Construct and Rationale of a Composite Performance Measure

A subcriterion specific for composite performance measures is included under Importance to Measure and Report (see 1d in Table 1). This is consistent with and refines the prior guidance regarding describing the purpose and quality construct for the composite performance measure. Composite measures are complex and represent a higher order construct than the individual measures. Justification for the approach to composite measure construction and analysis stems from the quality construct and rationale. Therefore, the quality construct should be clearly articulated and logical in order to meet this subcriterion and the must-pass criterion of Importance to Measure and Report.

Although the TEP recommended that the rationale for the composite performance measure be identified, it acknowledged that NQF endorses performance measures intended for both accountability and performance improvement and does not endorse measures for a specific accountability application (e.g., payment vs. public reporting). One TEP member suggested that the rationale should include the intended decision-making context (e.g., to select a provider for services, select a provider for contracting

or payment incentives, to identify or direct resources for improvement). While others noted that it might be difficult to envision how or why the component measures or the composite construction methodology should differ despite the decision-making context (given that all the decisions involve distinguishing good from poor quality), they did agree that measure developers should clearly explain how the aggregation and weighting of the components are consistent with the stated quality construct and rationale for the composite measure, including any decision-making context. The TEP acknowledged that endorsing multiple composite measures for a quality construct (such as quality of care for patients with congestive heart failure) that are created for different decision-making motivations could increase confusion and issues with competing measures. The decision-making context may influence whether a composite measure is useful at all. For example, a composite performance measure that includes multiple surgical mortality measures may be useful for assessing overall surgical quality, whereas the individual performance measures are more useful for selecting a hospital for a specific surgical procedure.

Scientific Acceptability of Measure Properties

Reliability

One cited advantage of composite performance measures is that using multiple indicators (components) increases reliability (i.e., the ability to detect a provider effect).[5] The purpose of combining individual measures that assess the quality of care provided to patients by providers or institutions is to determine whether these measures are useful in detecting a consistent pattern of practice or quality of care across patients of the provider or institution. That is, does a set of measures that, taken together, are thought to reflect good quality of care, show a more consistent pattern within a provider's practice or within an institution, and greater differences between providers or institutions than would be expected by chance alone? Reliability testing of the composite performance measure should demonstrate that the composite measure score differentiates signal from noise (i.e., random measurement error). It should be noted that increased reliability with increased number of indicators does not hold for all-or-none measures, when multiple indicators are essentially reduced to one data point.[8] Nevertheless, all-or-none performance measures also should demonstrate signal-to-noise reliability. These examples of measure reliability are different than a rationale that all-or-none measures are intended to foster a system perspective of care, sometimes called “system reliability.”

Although ideal, demonstrated signal-to-noise reliability of the individual component measures is not essential for having a reliable composite measure. In some cases, an individual performance measure may not provide a reliable signal because of small volume or rare events. However, that measure could appropriately be used as a component in a reliable composite performance measure.

Validity

Validity testing of the constructed composite performance measure score is more important than validity testing of the component measures because even if the individual component measures are valid, the aggregation and weighting rules for constructing the composite could result in a score that is not an accurate reflection of quality. However, some TEP members thought that requiring validity testing of the composite as a whole would be difficult to accomplish prior to NQF endorsement, although others questioned why NQF would endorse a performance measure without empirical evidence of validity. If validity of the composite performance measure is not demonstrated, then each of

the individual component measures must meet the NQF criteria for validity; further, validity testing of the overall composite measure would be expected by the time of endorsement maintenance.

It may be unlikely that another valid measure of the same quality construct (i.e., a criterion measure) will be available to test the criterion validity of a composite performance measure. Therefore, validity testing will require understanding and testing of various theoretical relationships. For example, a composite measure that includes multiple process measures could be tested for its association with a measure of a desired outcome. Alternatively, a composite measure might be tested for its ability to predict future outcomes or its ability to differentiate performance between groups known to differ on the particular quality construct.

Additional Testing of the Composite Performance Measure

A subcriterion specific for composite performance measures is included under Scientific Acceptability of Measure Properties (see 2d in Table 1). This is consistent with and refines the prior guidance regarding additional analyses to justify the construction of the composite measure (both component selection and aggregation and weighting rules). The initial criteria for testing were more relevant to composite measures that are based on correlated components. The modified criterion is neutral in terms of the analyses required. For example, if the rationale for summarizing the component measures in a composite is based on their correlation with each other, then analyses based on correlation (e.g., factor analysis, item-toal correlation, and inter-item correlation) are appropriate. In such cases, very high correlations between component measures may suggest that a component is redundant and not necessary. Conversely, ifthe rationale for summarizing the measures in a composite is not based on their correlation with each other, then analyses demonstrating the contribution of each component to the composite score, or their clinical justification (e.g., correlation of the individual component measures to a common outcome measure) are indicated.

The unit of analysis for which performance measures are calculated is typically the provider or institution (hospital, clinic, etc.) rather than the individual patient. For such measures, correlational analysis such as factor analysis or internal consistency reliability should be calculated at the level of the unit rather than patient, because the unit scores are what will be reported and acted upon. Correlations at the unit level might be quite different from those at the patient level. For example, in a patient survey, some respondents might tend to give more positive (or more negative) responses across the board, creating positive correlations among items that measure entirely distinct aspects of quality. However, when data are aggregated to the provider level, these patient tendencies average out, revealing correlations among items related at the provider level. As another example, measures of cardiac surgery might include complication rates during CABG surgery, during valve repair surgery, and during valve replacement surgery; since typically any patient undergoes only one of these procedures, the patient-level correlations of these measures are not defined but correlations at the provider or hospital level are meaningful and could be examined to assess the validity of a composite surgical quality measure. However, special statistical methods should be used for estimating such unit-level correlations, especially when the component measures do not have high unit-level reliability.

Composite measures are, by definition, complex; however, secondary objectives for composite measure construction are parsimony regarding the component selection and simplicity in terms of the aggregation and weighting. Scientific Acceptability of Measure Properties is a must-pass criterion and

measures must meet both reliability and validity. In addition, composite measures must meet this additional criterion in order to meet the must-pass criterion of Scientific Acceptability.

Feasibility

The standard feasibility criteria apply to the composite measure as a whole, but must take into account all the component measures. That is, feasibility of the composite measure will be influenced by the least feasible of the component measures.

Usability and Use

Composite performance measures must meet the updated criteria for Usability and Use. The TEP noted that disaggregation of a composite measure is not an absolute requirement because the individual component measures need not be independently reliable. However, at a minimum, the components of the composite performance measure must be identified. For purposes of improvement, the data must be collected so as to facilitate investigation of the individual components.

Comparison to Related and Competing Measures

Composite performance measures are subject to comparison to related and competing measures. If the component measures are not NQF-endorsed, they must be harmonized with endorsed measures or assessed against competing measures.

Table 1. NQF Measure evaluation Criteria and Guidance for Evaluating Composite Performance Measures

Measure Evaluation Criteria	**Guidance for Composite Performance Measures**
Conditions	
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.	The evidence criterion (1a) must be met for each component of the composite (unless NQF-endorsed under the current evidence requirements).

Measure Evaluation Criteria	Guidance for Composite Performance Measures

• 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).
 	
Composite 1d. For composite performance measures, the following must be clearly stated and logical:
• The quality construct, including the representativeness of the component measures and the relationship of the component measures to the composite and to each other; and
• The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
• How the aggregation and weighting of the component measures are consistent with and representative of the stated quality construct and rationale.
	The performance gap criterion (1b) must be met for each component (unless NQF-endorsed), 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.
	
	Subcriterion 1d must be met for a composite performance measure to meet the criterion of Importance to Measure and Report.
	All three elements must be clearly articulated and represent a logical rationale.

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*	

Measure Evaluation Criteria	Guidance for Composite Performance Measures
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 allows for comparability. EHR measure specifications are based on the quality data model (QDM). [9] **Add to Note 8:** Composite measure specifications include scoring rules (i.e., how the component scores are combined or aggregated), how missing data are handled, required sample sizes; and when appropriate methods for standardizing scales across component scores and weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite) **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	2a2. For composite performance measures, reliability must be demonstrated for the composite measure score. Reliability of the individual component measures is not sufficient, and in some cases, component measures that are not independently reliable can contribute to reliability of the composite measure. However, if the component measures will be disaggregated in accountability applications, then reliability for the component measures must be demonstrated (unless NQF-endorsed). 2b2. For composite performance measures, validity should be demonstrated for the composite measure score. If not feasible at the time of initial endorsement, validity of the component measures must meet NQF criteria, and by endorsement maintenance, validity of the composite performance measure must be demonstrated. If the component measures will be disaggregated for accountability applications, then validity for the component measures must be demonstrated (unless NQF-endorsed). 2b3. Exclusions apply primarily to the component measures and would not

Measure Evaluation Criteria	Guidance for Composite Performance Measures
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]	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 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, socioeconomic status, gender); **OR** rationale/data justifies why stratification is not necessary or not feasible.	2c. Applies to composite performance measures.
Composite 2d. For composite performance measures, empirical analyses demonstrate: • the component measures fit the quality construct and are necessary (secondary objective of parsimony to the extent possible); • the aggregation and weighting rules are consistent with the quality construct and rationale (secondary objective of simplicity to the extent possible); and • the extent of missing data and how the specified handling of missing data minimizes bias.	Subcriterion 2d must be met for a composite performance measure to meet the criterion of Scientific Acceptability of Measure Properties.
3. Feasibility: Extent to which the required data are readily available or	

Measure Evaluation Criteria	Guidance for Composite Performance Measures
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 as a whole, taking into account all component measures.
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).	4a. Applies to composite performance measures. To facilitate transparency, at a minimum, the individual components of the composite must be identified; ideally, if the components are NQF-endorsed and/or meet reliability and validity criteria, the component scores would be reported. 4b. Applies to composite performance measures, except it should explicitly link back to the quality construct and rationale. To facilitate improvement, data should be collected in such a way as to permit disaggregation. 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.

Measure Evaluation Criteria	Guidance for Composite Performance Measures
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.	5a and 5b. Applies to composite performance measures as a whole as well as the component measures.

1

2 Table 2. Notes to Measure Evaluation Criteria

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

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. Composite measure specifications include scoring rules (i.e., how the component scores are combined or aggregated), how missing data are handled, required sample sizes; and when appropriate methods for standardizing scales across component scores and weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite).
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.

Feasibility

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.
Usability and Use
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.
Comparison to Related and Competing Measures
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.

Recommendations for Review Process

The TEP made several recommendations for the process of evaluating composite performance measures.

- 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.
- If a steering committee recommends the removal of one or more components from the composite performance measure—and the developer is agreeable to the revised construction of the

composite—there should be an opportunity for the developer to respond to the recommendation within the project rather than having to completely re-submit the revised measure at a later date.

- Provide examples of types of analyses for different types of composite performance measures. (See Appendix B for a first step.)

Notes

1. Institute of Medicine, *Performance Measurement: Accelerating Improvement,* Washington, DC: National Academies Press; 2006.
2. National Quality Forum (NQF), *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety-Composite Measures: A Consensus Report,* Washington, DC: National Quality Forum; 2009.
3. Booysen F, An overview and evaluation of composite indices of development, *Social Indicators Research,* 2002;59:115-151.
4. Fayers PM, Hand DJ, Causal variables, indicator variables and measurement scales: an example from quality of life, *J R Statist Soc A,* 2002;165 (Part 2):233-261.
5. Kaplan SH, Normand SL, ., *Conceptual and Analytical Issues in Creating Composite Measures of Ambulatory Care Performance,* Washington, DC: National Quality Forum; 2006.
6. Nardo M, Saisana M, Saltelli A, et al., *Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Statistics Working Paper,* Paris, France: OECD Statistics Directorate; 2005. Report No.: STD/DOC(2005)3.
7. O'Brien SM, Shahian DM, DeLong ER, et al., Quality measurement in adult cardiac surgery: part 2--Statistical considerations in composite measure scoring and provider rating, *Ann Thorac Surg,* 2007;83(4 Suppl):S13-S26.
8. Reeves D, Campbell SM, Adams J, et al., Combining multiple indicators of clinical quality: an evaluation of different analytic approaches, *Med Care,* 2007;45(6):489-496.
9. Shahian DM, Edwards FH, Ferraris VA, et al., Quality measurement in adult cardiac surgery: part 1--Conceptual framework and measure selection, *Ann Thorac Surg,* 2007;83(4 Suppl):S3-12.

Appendix A: 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	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, 2007
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

Term	Definition	Source
		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 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	NQF, Composite Guidance Report, 2007, Aligning Forces, 2010, Reeves, 2007

Term	Definition	Source
	patients= 100+75 / 2. (Has also been called “patient average”.)	
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
Performance measure	Numeric quantification of healthcare quality for a designated accountable healthcare entity, such as hospital, health plan, nursing home, clinician, etc.	PRO Report, 2012
Psychometric approach	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
Quality construct	A hypothetical complex concept of quality.	
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

1

Appendix B: Approaches for Constructing Composite Performance Measures

Quality Construct	Description	Unique Considerations for Testing and Evaluating the Composite
1. The quality construct is seen as causing the component performance measure scores • Also known as psychometric, reflective, scale, homogenous scale, dimensional • Example: NQF# 0530: Mortality for Selected Conditions (AHRQ)	• Scores on the component performance measures are considered the effect of quality (or caused by quality) • Component performance measures are considered a random sample of potential indicators of quality and should be interchangeable; therefore, focusing QI only on the component performance measures may not change the composite score • Component performance measures should be correlated with one another because they share common variance; and each component is correlated with the total composite score (omitting the component being assessed) **Aggregation:** **Combination of multiple individual performance measures** Various approaches may be used, including: ▪ Opportunities [sum of all numerators / sum of all denominators] ▪ Average/weighted average of component measure scores [score on A + score on B + score on C . . . / # of component performance measures]; or ▪ Comparison to some benchmark (e.g., percentage of component performance measures that improved, reached 80%, etc.)	
2. The quality construct is seen as being caused or defined by the component performance measure scores • Also known as clinimetric, formative,	• Component performance measures are considered to cause quality (or define quality) • Component performance measures define the quality construct and should cover the entire scope of the quality construct; therefore, focusing QI on the component performance measures should change the composite score • Component performance measures do not	

Quality Construct	Description	Unique Considerations for Testing and Evaluating the Composite
index, heterogenous index, categorical Example:	need to be correlated with one another (but could be);?? each component should be correlated with the total composite score (omitting the component being assessed)?? **Aggregation:** **Combination of multiple individual performance measures** Various approaches may be used, including: ▪ Opportunities [sum of all numerators / sum of all denominators] ▪ Average/weighted average of component measure scores [score on A + score on B + score on C . . . / # of component performance measures]; or ▪ Comparison to some benchmark (e.g., percentage of component performance measures that improved, reached 80%, etc.)	
3. The quality construct is viewed as receiving all necessary care • Also known as All-or-None Example: NQF# 0729: Optimal Diabetes Care (MN Community Measurement)	Individual patient scores on component measures are considered to define quality and ALL must be achieved to signal quality **Aggregation:** **Composite numerator** - Multiple components specified in the numerator and measured for each patient Percentage of patients who received all necessary components of care [# of patients in the denominator who met all components (A **and** B **and** C **and** . . .) / # of patients in target population]	
4. The quality construct is viewed as receiving necessary care, but receiving some is better	Individual patient scores on component measures are considered to define quality and achieving more is a signal of better quality	

Quality Construct	Description	Unique Considerations for Testing and Evaluating the Composite
than none • Also known as partial credit, percentage of necessary care Example:	**Aggregation**: **Composite numerator** - Multiple components specified in the numerator and measured for each patient Average percentage of necessary components of care received by patient [Sum of percentage of components met (A, B, C . . .) for each patient in the denominator / # of patients in target population]	
5. The quality construct is viewed as not experiencing any healthcare-acquired adverse event/complication • Also known as any-or-all Example: NQF# 0564: Complications within 30 Days Following Cataract Surgery Requiring Additional Surgical Procedures (PCPI)	Individual patient scores on component measures are considered to define quality and NONE must be achieved to signal quality **Aggregation:** **Composite numerator** - Multiple components specified in the numerator and measured for each patient Percentage of patients who experienced any of the component adverse events or complications [# of patient in the denominator who experienced A **or** B **or** C **or** . . . . / # of patients in target population]	
6. The quality construct is defined by one concept but uses additional information on average performance to increase precision (reliability) Also known as reliability adjustment, shrinkage estimator	• Quality is defined by one performance measure but this measure is considered an unreliable estimate by itself, because of rare events or small case volume **Aggregation:** • Combines two rates of the same concept (e.g., a provider's observed mortality rate and an average mortality rate for a specific category of providers such as quartile by patient case volume) • To-date has been used only with outcome measures	

Quality Construct	Description	Unique Considerations for Testing and Evaluating the Composite
Example: NQF# 0737: Survival Predictor for Esophagectomy Surgery (Leapfrog)	• Uses a provider characteristic (e.g., case volume)to categorize all providers for purposes of creating average rates for different categories (Weight x observed rate) + (weight x average rate) Weight is based on reliability of the provider observed rate, which is influenced by case volume	

1

2 **The following are examples are not composite performance measures**

Conceptual Model	Description	Unique Considerations for evaluation
7. Multi-item composites to measure individuals regardless of quality construct Example: Model 1-PHQ-9, CAHPs; Model 2-Apgar	• Multi-item scale, instrument, index, survey administered to individuals. • Patient data on these scales may be used in an individual performance measure or a composite performance measure; but the scale itself is not a performance measure and not eligible by itself for NQF endorsement.	• Not a composite *performance* measure • If patient data from such a scale is used in a performance measure, the reliability and validity of the scale also must be demonstrated. • See PRO project.
8. Multiple aspects of quality are identified, but no quality construct for a composite is provided Example: NQF# 0101 Falls: Screening, Risk-Assessment, and Plan of Care to Prevent Future Falls (NCQA) CAHPS performance measures	The performance measures represent multiple individual aspects of quality There are two variants of this example: • Individual performance measures are identified as paired or grouped to be used and reported together to appropriately interpret results • Multiple related individual performance measures are submitted on one submission form, but require computation of individual performance measure scores; may have multiple denominators as well as	• Identifying paired or grouped measures is appropriate in some circumstances; however, individual performance measures need to be evaluated individually against all criteria and should be submitted on separate forms. • Including multiple individual measures in one form could obscure measure-specific information, making evaluation more difficult; it also is more difficult for others to find performance measures when they are catalogued with one

	numerators. They do not necessarily need to be reported together for appropriate interpretation and if that were the case could be submitted as paired/grouped measures.	measure number.

1

2

Appendix C: References Consulted

1. Agency for Healthcare Research and Quality (AHRQ), *Inpatient Quality Indicators Composite Measure Workgroup Final Report,* 2008. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/PSI%20Composite%20Development.pdf. Last accessed October 2012.

2. Asch S, Hofer T. Representing overall quality of care: The whole must be more than the sum of the parts. 2008.

3. Ashby J, Juarez DT, Berthiaume J, et al., The relationship of hospital quality and cost per case in Hawaii, *Inquiry,* 2012;49(1):65-74.

4. Booysen F, An overview and evaluation of composite indices of development, *Social Indicators Research,* 2002;59:115-151.

5. Diamantopoulos A, Winklhofer HM, Index construction with formative indicators: An alternative to scale development, *Journal of Marketing Research,* 2001;38(2):269-277.

6. Dijkers MP, Psychometrics and clinimetrics in assessing environments. A comment suggested by Mackenzie et al., 2002, *J Allied Health,* 2003;32(1):38-43.

7. Dimick JB, Staiger DO, Baser O, et al., Composite measures for predicting surgical mortality in the hospital, *Health Aff (Millwood),* 2009;28(4):1189-1198.

8. Dimick JB, Staiger DO, Osborne NH, et al., Composite measures for rating hospital quality with major surgery, *Health Serv Res,* 2012;47(5):1861-1879.

9. Eapen ZJ, Fonarow GC, Dai D, et al., Comparison of composite measure methodologies for rewarding quality of care: an analysis from the American Heart Association's Get With The Guidelines program, *Circ Cardiovasc Qual Outcomes,* 2011;4(6):610-618.

10. Edwards JR, Bagozzi RP, On the nature and direction of relationships between constructs and measures, *Psychol Methods,* 2000;5(2):155-174.

11. Fayers PM, Hand DJ, Causal variables, indicator variables and measurement scales: an example from quality of life, *J R Statist Soc A,* 2002;165 (Part 2):233-261.

12. Felt-Lisk S, Lavin B, Gold M, Aggregate quality measures for the National Healthcare Quality Report: Summary of technical advisory panel meetings May/June 2005 (DRAFT), 2005;

13. Hess BJ, Weng W, Lynn LA, et al., Setting a fair performance standard for physicians' quality of patient care, *J Gen Intern Med,* 2011;26(5):467-473.

14. Holmboe ES, Weng W, Arnold GK, et al., The comprehensive care project: measuring physician performance in ambulatory practice, *Health Serv Res,* 2010;45(6 Pt 2):1912-1933.

15. Ingenix, Creating quality composite scores: Challenges and issues in physician quality measurement, 2008;October 2012.

16. Institute of Medicine, *Performance Measurement: Accelerating Improvement,* Washington, DC: National Academies Press; 2006.

17. Jacobs R, Goddard M, Smith PC, How robust are hospital ranks based on composite performance measures?, *Med Care,* 2005;43(12):1177-1184.

18. Jacobs R, Goddard M, Smith PC, *Public Services: Are Composite Measures a Robust Reflection of Performance in the Public Sector,* 2006. Report No.: CHE Research Paper 16.

19. Kaplan SH, Griffith JL, Price LL, et al., Improving the reliability of physician performance assessment: identifying the "physician effect" on quality and creating composite measures, *Med Care,* 2009;47(4):378-387.

20. Kaplan SH, Normand SL, *Conceptual and Analytical Issues in Creating Composite Measures of Ambulatory Care Performance,* Washington, DC: National Quality Forum; 2006.

21. Kianifard F, Evaluation of clinimetric scales: Basic principles and methods, *The Statistician,* 1994;43(4):475-482.

22. Nardo M, Saisana M, Saltelli A, et al., *Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Statistics Working Paper,* Paris, France: OECD Statistics Directorate; 2005. Report No.: STD/DOC(2005)3.

23. National Committee for Quality Assurance (NCQA), MEMO: Summary of Alliance Use of Composite Measures, 2010;

24. National Quality Forum (NQF), *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety-Composite Measures: A Consensus Report,* Washington, DC: National Quality Forum; 2009.

25. Nolan T, Berwick DM, All-or-none measurement raises the bar on performance, *JAMA,* 2006;295(10):1168-1170.

26. Normand C, Measuring outcomes in palliative care: limitations of QALYs and the road to PALYs, *J Pain Symptom Manage,* 2009;38(1):27-31.

27. O'Brien SM, Shahian DM, DeLong ER, et al., Quality measurement in adult cardiac surgery: part 2--Statistical considerations in composite measure scoring and provider rating, *Ann Thorac Surg,* 2007;83(4 Suppl):S13-S26.

28. Peterson ED, DeLong ER, Masoudi FA, et al., ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment: a report of American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to Develop a Position Statement on Composite Measures), *J Am Coll Cardiol,* 2010;55(16):1755-1766.

29. Peterson ED, DeLong ER, Masoudi FA, et al., ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment: a report of the American College of Cardiology

Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to develop a position statement on composite measures), *Circulation,* 2010;121(15):1780-1791.

30. Physician Consortium for Performance Improvement (PCPI), *Measures Development, Methodology, and Oversight Advisory Committee: Recommendations to PCPI Work Groups on Composite Measures,* 2010. Available at http://www.ama-assn.org/resources/doc/cqi/composite-measures-framework.pdf. Last accessed October 2012.

31. Reeves D, Campbell SM, Adams J, et al., Combining multiple indicators of clinical quality: an evaluation of different analytic approaches, *Med Care,* 2007;45(6):489-496.

32. Ross JS, Correlation of inpatient and outpatient measures of stroke care quality within Veterans Health Administration hospitals, 2011;42(8):2269-2275.

33. Scholle SH, Roski J, Adams JL, et al., Benchmarking physician performance: reliability of individual and composite measures, *Am J Manag Care,* 2008;14(12):833-838.

34. Scholle SH, Roski J, Adams JL, et al., Benchmarking Physician Performance: Reliability of Individual and Composite Measures, *American Journal of Managed Care,* 2008;14(12):829-838.

35. Shahian DM, Edwards FH, Ferraris VA, et al., Quality measurement in adult cardiac surgery: part 1--Conceptual framework and measure selection, *Ann Thorac Surg,* 2007;83(4 Suppl):S3-12.

36. Shwartz M, Ren J, Pekoz EA, et al., Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights, *Med Care,* 2008;46(8):778-785.

37. Staiger DO, Dimick JB, Baser O, et al., Empirically derived composite measures of surgical performance, *Med Care,* 2009;47(2):226-233.

38. Streiner DL, Clinimetrics vs. psychometrics: an unnecessary distinction, *J Clin Epidemiol,* 2003;56(12):1142-1145.

39. Streiner DL, Being inconsistent about consistency: when coefficient alpha does and doesn't matter, *J Pers Assess,* 2003;80(3):217-222.

40. Timbie JW, Shahian DM, Newhouse JP, et al., Composite measures for hospital quality using quality-adjusted life years, *Stat Med,* 2009;28(8):1238-1254.

41. Weifeng W, Hess BJ, Lynn LA, et al., Measuring physicians' performance in clinical practice: reliability, classification accuracy, and validity, *Eval Health Prof,* 2010;33(3):302-320.

Appendix D: Technical Expert Panel and NQF Staff

TECHNICAL EXPERT PANEL

Patrick Romano, MD, MPH (Co-Chair)
UC Davis School of Medicine
Sacramento, CA

Elizabeth R. DeLong, PhD (Co-Chair)
Duke University Medical Center
Durham, NC, State

John D. Birkmeyer, MD
University of Michigan
Ann Arbor, MI, State

Dale Bratzler, DO, MPH
Oklahoma University Health Services Center
Oklahoma City, OK, State

James Chase, DO, MPH
Minnesota Community Measurement
Minneapolis, MN, State

Nancy Dunton, PhD, FAAN
University of Kansas Medical Center, School of Nursing
Overland Park, KS, State

Elizabeth Goldstein, PhD
Centers for Medicare and Medicaid Services
Baltimore, MD, State

Sherrie Kaplan, PhD, MPH
The University of California - Irvine
Irvine, CA, State

Lyn Paget, MPH
Informed Medical Decisions Foundation
Boston, MA, State

David Shahian, MD
Massachusetts General Hospital
Boston, MA, State

Steven Wright, PhD
Veteran's Health Administration
Providence, RI, State

Alan Zaslavsky, PhD
Harvard Medical School
Boston, MA, State

NQF STAFF

Helen Burstin, MD, MPH
Senior Vice President
Performance Measures

Heidi Bossley, MSN, MBA
Vice President
Performance Measures

Karen Pace, PhD, RN
Senior Director
Performance Measures

Karen Johnson, MS
Senior Director
Performance Measures

Elisa Munthali, MPH
Senior Project Manager
Performance Measures