

Composite Performance Measure Evaluation Guidance

April 8, 2013



NATIONAL
QUALITY FORUM

Contents

- Introduction 1
 - Purpose 1
- Background 2
 - Prior Guidance on Evaluating Composite Measures..... 2
 - NQF Experience with Composite Performance Measures..... 2
- Definition of Composite Performance Measure 4
 - Types of Composite Performance Measures 4
 - Key Steps in Developing a Composite Performance Measure..... 7
- Guiding Principles 7
 - Terminology 7
 - Component Measures..... 7
 - Composite Performance Measure 8
- Recommendations for Composite Performance Measure Evaluation 9
 - Importance to Measure and Report 10
 - Scientific Acceptability of Measure Properties..... 12
 - Feasibility 14
 - Usability and Use 14
 - Comparison to Related and Competing Measures 15
- Recommendations for Review Process..... 23
- Next Steps 24
- Appendix A: Glossary 27
- Appendix B: Approaches for Constructing Composite Performance Measures..... 31
- Appendix C: References Consulted 35
- Appendix D: Technical Expert Panel and NQF Staff 39

Composite Performance Measure Evaluation Guidance

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 performance 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 informed decisionmaking by multiple stakeholders, including consumers, purchasers, and policy makers.

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

NQF endorses *performance measures* that are intended for use in both performance improvement and accountability applications. Several composite measures are included in NQF's portfolio of endorsed measures, and NQF previously developed guidance² to assist NQF 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 reviewing papers describing a wide variety of evidence and experience around composite measures, and 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 concept 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, 31 measures submitted to NQF for potential endorsement have been flagged as composite measures. Of these, 25 are currently endorsed.^a Many 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=5) and composites constructed using various methods of aggregation and weighting methodologies (n=9). 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 all of the required processes); simpler all-or-none measures requiring that linked multiple conditions are met (e.g., assess vaccination status and administer flu vaccine); and any-or-none 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

^a See NQF's [Quality Positioning System](#) search tool; for a list of currently endorsed composite measures, filter based on measure type.

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 subcriterion is whether measures of processes that are distal to desired outcomes could be included in composite measures. For example, a performance measure of merely obtaining a lab test is often considered not to meet the evidence subcriterion because it is so distal to the desired outcome and is usually based on expert opinion rather than direct evidence; 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 criterion 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, for a performance measure such as 90-minute door-to-balloon time for cardiac catheterization, a balancing measure might be the rate of premature or unwarranted activation of the catheterization lab team, which is an undesirable and costly unintended negative consequence.

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 (from multiple developers) are 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, as noted above. In such cases, these components either compete directly with other endorsed measures or are 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 have been 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, NQF 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

The TEP reviewed and retained the definition provided in the initial composite report and added explicit clarification that it refers to composite *performance* measures. **A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.**

While the term “composite measure” also has been used to refer to multi-item instruments and scales used to obtain data from individuals about a particular domain of health status, quality of life, or experience with care (e.g., CAHPS; PHQ-9), such instruments or scales alone do not constitute a *performance* measure—either individual or composite. However, if considered a reflection of performance, aggregated data from multi-item instruments or scales can be used as the basis of an individual performance measure, which can, in turn, be used as a component of a composite performance measure. Note that use of patient-reported outcomes (PROs) in performance measurement is the subject of a recent NQF project (see [PRO report](#)).

Types of Composite Performance Measures

The TEP acknowledged that a simple definition is not sufficient for determining whether a performance measure should be considered a composite for purposes of NQF measure submission, evaluation, and endorsement. The TEP considered various ways to provide further guidance.

Composites often are classified according to the empirical and conceptual relationships among the component measures and between the components and the composite, the approaches 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). The glossary in [Appendix A](#) contains definitions for various approaches to combining the component measures. [Appendix B](#) provides a description of various models of the relationships among component measures and with the overall composite.

The TEP decided that a formal classification of types of composite performance measures or models would not be particularly useful and could lead to unnecessary attention to naming the approach used to construct the composite. Regardless of the approach to constructing the composite, a coherent quality construct and rationale should guide composite development, testing, analysis, and evaluation. The TEP 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.

However, to ensure that developers know expectations in advance and submit the additional information and analyses needed to support evaluation of a composite performance measure based on the criteria and guidance provided in this report, it is necessary to clearly delineate what types of

measures will be considered to be composite performance measures for the purposes of NQF measure submission, evaluation, and endorsement (Box 1).

Box 1. Identification of Composite Performance Measures for Purposes of NQF Measure Submission, Evaluation, and Endorsement*

The following **will be** considered composite performance measures for purposes of NQF endorsement:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures **assessed separately for each patient** and then aggregated into one score for an accountable entity. These include:
 - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
 - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

The following **will not be** considered composite performance for purposes of NQF endorsement at this time:

- Single performance measures, even if the data are patient scores from a composite instrument or scale (e.g., single performance measure on communication with doctors, computed as the percentage of patients where the average score for four survey questions about communication with doctors is equal or greater than 3).
- Measures with multiple measure components that are assessed for each patient, but that result in multiple scores for an accountable entity, rather than a single score. These generally should be submitted as separate measures and indicated as paired/grouped measures.
- Measures of multiple linked steps in one care process assessed for each patient. These measures focus on one care process (e.g., influenza immunization) but may include multiple steps (e.g., assess immunization status, counsel patient, and administer vaccination). These are distinguished from all-or-none composites that capture multiple care processes or outcomes (e.g., foot care, eye care, glucose control).
- Performance measures of one concept (e.g., mortality) specified with a statistical method or adjustment (e.g., empirical Bayes shrinkage estimation) that combines information from the accountable entity with information on average performance of all entities or a specified group of entities (e.g., by case volume), typically in order to increase reliability.

* The list in Box 1 includes the types of measure construction most commonly referred to as composites, but this list is not exhaustive. NQF staff will review any potential composites that do not clearly fit one of these descriptions and make the determination of whether the measure will be evaluated against the additional criteria for composite performance measures.

Discussion

The key feature of a composite is combining information from multiple measures into a single score; therefore, single measures or any measure that results in multiple scores are by definition not composites for purposes of NQF endorsement. The TEP was not in complete agreement regarding the identification of composites for purposes of NQF measure submission, evaluation, and endorsement.

Therefore, the decisions regarding identification of composite performance measures listed in Box 1 should be reviewed again after gaining more experience with composites.

Measures with two or more individual performance measure scores combined into one score for an accountable entity. The TEP agreed that these are composite performance measures.

Measures with two or more individual component measures *assessed separately for each patient and then aggregated into one score for an accountable entity.* All-or-none measures assess whether all essential care processes were received, or all outcomes were experienced, by each patient. Any-or-none measures assess whether any of a list of adverse outcomes were experienced, or inappropriate or unnecessary care processes were received,^b by each patient. Although not unanimous, a majority of the TEP agreed that all-or-none and any-or-none measures should be considered composite performance measures. These measures are similar in construction in that all the components are assessed separately for each patient.

The TEP discussed whether any-or-none measures that include a group of patient-specific outcomes, such as complications, should always be considered composites. For surgical patients, for example, a developer may create a measure that looks for various events that may occur as unintended consequences of the operation for each patient. These measures may include events that have not previously been considered as individual measures (e.g., hemorrhage) or events that have previously been considered as individual measures (e.g., death, readmission). In some instances, the developer may not view a measure that incorporates multiple events such as complications as an any-or-none composite (e.g., complications are viewed as a single measure instead of multiple measures). The CSAC agreed that such measures will be considered composites, with the expectation that the information needed to evaluate the composite-specific criteria is provided. However, if the developer provides a conceptual justification as to why such a measure should not be considered a composite, and that justification is accepted by the NQF steering committee, the measure can then be considered a single measure rather than a composite.

Measures of multiple linked steps in one care process assessed for each patient. The TEP considered whether measures that focus on one care process (e.g. influenza immunization) but include multiple linked steps (e.g., assess immunization status, counsel patient, and administer vaccination) assessed for each patient should be considered as composites. Typically, these types of measures have not been submitted as composites; however they are constructed similarly to all-or-none measures. Usually the evidence for such measures is specific to the treatment step, in contrast to more typical all-or-none measures that focus on multiple care processes (e.g., foot care, eye care, glucose control). The majority of TEP members did not view these types of measures as composites; therefore, they will not be considered composites for purposes of NQF measure submission, evaluation, and endorsement.

Performance measures of one concept specified with a statistical method or adjustment that combines information from the accountable entity with information on average performance of all entities or a specified group of entities. Some performance measures use statistical methods such as empirical Bayes shrinkage to obtain more reliable estimates (e.g., mortality, readmission). These

^b Any-or-none measures could conceivably include a group of treatments or services that patients should not receive to address overuse or other types of inappropriate care; however, currently there are no examples of this in the NQF portfolio.

statistical techniques combine information on each individual entity’s performance with information on the average performance of all entities or a grouping of entities on some characteristic (e.g., by case volume). To date, most of the measures submitted to NQF that have utilized this approach have combined the provider-specific results with the overall average performance. Others have combined the provider-specific results with the average performance for their particular volume category. Some have referred to these measures as composites, others have referred to this as applying a reliability adjustment, and others just name the statistical model (i.e., Bayesian hierarchical modeling). The majority of TEP members did not view these types of measures as composites; therefore, they will not be considered composites for purposes of NQF measure submission, evaluation, and endorsement. The current NQF criteria for risk adjustment of outcomes, reliability, and validity are appropriate for evaluating these types of measures.

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 the rationale for the composite (including its putative advantages compared with relevant individual measures);
- 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 and guidance regarding the evaluation criteria.

Terminology

- As noted above, the TEP opted for a broad, generic definition of composite performance measure.
- The term “composite measure” is used in reference to individual-level instruments or scales as well as aggregate-level performance measures. NQF only endorses *performance measures*, not the individual-level instrument or scale.
- 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 and not necessarily constrained to adhere to a specific method or categorization (e.g., “psychometric” and “clinimetric”).

Component Measures

The prior composite evaluation criteria required that each component 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 combining multiple measures 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 the relevant information from that endorsement process could be referenced during review of the composite performance measure.

- 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). In some cases, the evidence may be for a group of interventions included in a composite performance measure, rather than each one separately.
- NQF-endorsement of the individual component measures should not be mandatory; however, NQF endorsement of component performance measures could satisfy some requirements for those component measures that are included in a composite (e.g., a developer would not have to demonstrate the reliability/validity of a component measure that is currently endorsed).
- The individual components in a composite performance measure generally should demonstrate a gap in performance; however, there may be conceptual (e.g., clinical evidence) or analytical justification (e.g., addition increases the variability/gap for the overall composite measure) 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. As several authors have noted, “a construct... can be viewed as the cause of individual quality indicators; that is, the quality indicators reflect or manifest the extent to which the organization has achieved quality... Alternatively, the construct can be viewed as formed from the indicators... The first type of relationship is called reflective and the second formative... A reflective construct assumes that causality flows from the construct to the indicators, while in a formative model, causality flows from the indicators to the construct.”¹⁰ A construct has been defined as “a conceptual term used to describe a phenomenon of theoretical interest... The phenomena that constructs describe can be unobservable (e.g., attitudes) or observable (e.g., task performance)... In either case the construct itself is an abstract term.”¹¹ In this report, the term “quality construct” is applied both to “reflective composites” and to “formative composites”.

Component measures should be selected based on fit with the quality construct, and analyses should support that fit. All composite performance measures share the potential for simplification by presenting one summary score instead of multiple scores for individual performance measures. However, simplification alone is not sufficient justification for a composite performance measure. Each component should fit the quality construct and provide added value to the composite. The composite performance measure should similarly provide added value relative to having individual performance measures. Composite measures are complex, with aggregation and weighting rules that do not apply to

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 guarantee 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.
- Each component of a composite performance measure should provide added value to the composite as a whole—either empirically (e.g., they contribute to the reliability, or overall score) or conceptually. A related objective is parsimony. However, having a complete set of component measures from all relevant performance domains may be conceptually preferable to dropping measures that do not empirically contribute to comparative evaluation of health care entities.
- 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 related objective is methodological simplicity. However, complex aggregation and weighting rules may improve the reliability and validity of a composite performance measure, relative to simpler aggregation and weighting rules.
- The standard NQF criteria apply to composite performance measures.
- NQF only endorses performance measures that are intended for use in both performance improvement and accountability applications.

Recommendations for Composite Performance Measure Evaluation

The NQF performance measure evaluation criteria apply to composite performance measures and their component measures. Evaluation of composite performance measures can, to a large extent, be incorporated into the standard NQF criteria and processes. NQF endorsement is not necessary 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 subcriteria are needed to evaluate composite performance measures; these are incorporated into the evaluation criteria in Table 1 (see [1d](#) and [2d](#)) and discussed below. The NQF measure submission form will be modified to accommodate the composite-specific subcriteria.

The TEP recommended that NQF Steering Committees include members with experience and expertise to evaluate composite performance measures against these criteria. The TEP also noted that composite methodology is still an evolving area of performance measurement that is varied and complex. Both the guidance for the identification of composite performance measures for purposes of NQF endorsement, as well as the recommended evaluation criteria, will need to be reviewed again in the future to determine if further refinements are needed.

Importance to Measure and Report

Evidence

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 appropriate 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 appropriate reflection of quality.

Each component measure must meet the evidence subcriterion 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 subcriterion. If any component measure does not meet the evidence subcriterion, or does not qualify for an exception to the evidence subcriterion, then the composite would not meet the criterion for Importance to Measure and Report unless those components were removed. Evidence is required for each component measure, regardless of the approach to constructing a composite measure (i.e., all-or-none, any-or-none, or combining scores from individual performance measures). The evidence may be for a group of interventions included in a composite performance measure, rather than for each one separately (if that is how they were studied). For example, in studies that include multiple interventions delivered to all subjects in the treatment group, the effect of each intervention on outcomes cannot be disaggregated.

Performance Gap

As with individual performance measures, effort for measurement should be directed to where there is variation or overall poor performance. Therefore, the composite performance measure as a whole should demonstrate a performance gap. Each component measure also should generally meet the criterion of performance gap to justify its inclusion in the composite. 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).

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 subcriterion is consistent with and refines the prior guidance

regarding a description of the purpose and quality construct for the composite performance measure. Quality of care is an abstract concept that is measured using observed variables. Composite measures are complex, multidimensional, and represent a higher order construct than individual measures. The composite performance measure quality construct is a concept of quality that includes:

- the overall area of quality (e.g., quality of CABG surgery);
- the included component measures (e.g., pre-operative beta blockade; CABG using internal mammary artery; CABG risk-adjusted operative mortality);
- the conceptual relationships between each component and the overall composite (e.g., components cause or define quality, components are caused by or reflect quality); and
- the relationships among the component measures (e.g., whether they are correlated or not, processes that are expected to lead to better outcomes).

The TEP agreed that the rationale for constructing a composite performance measure, including how the composite provides a distinctive or additive value over the component measures individually, should be described. The TEP 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). However, the TEP discussed that at times, the decisionmaking context could influence the composite measure construction, i.e., which component measures are included or the aggregation and weighting rules.^c The decision-making context also could influence whether a composite measure is more useful than individual performance measures.^d Additionally, multiple composite measures for the same or similar quality construct, even if addressing different decisionmaking motivations, will trigger an evaluation of competing measures and the rationales may be an important aspect of determining whether multiple endorsed composite performance measures are justified.

Some TEP members thought the decisionmaking context is a unique aspect of composite performance measures in that the appropriate aggregation and weighting of component measures may vary according to the intended use of the composite. Other TEP members expressed concern that identifying a specific decisionmaking context might be viewed as inconsistent with NQF's current policy to endorse measures suitable for both accountability and performance improvement, rather than for a specific accountability application. Some noted that all composites should be a reflection of quality and therefore the decisionmaking process should be irrelevant. However, the rationale for the composite could include the intended decisionmaking context (e.g., to select a provider for surgery, to create payment incentives to direct resources for improvement), if that context is relevant to explaining how the composite is constructed.

^c For example, hospital performance on two related sets of measures (A and B) may be important to patients, but failure on group A measures may entail additional costs to the hospital (e.g., longer mean LOS for Medicare fee-for-service patients) whereas failure on group B measures may not entail such additional costs. Composites intended to inform patient choice should include both sets of measures, whereas a pay-for-performance program might use a composite limited to B measures, because the hospital already has a financial incentive to improve on A measures, and therefore the financial reward should be targeted to stimulate improvement on B measures.

^d 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.

Justification for the inclusion of particular component measures and the approach to composite measure construction and analysis stems from the quality construct and rationale. Therefore, the quality construct and rationale should be clearly articulated and logical in order to meet this subcriterion (1d). Importance to Measure and Report is a must-pass criterion and composite performance measures must meet all subcriteria, including 1d.

Scientific Acceptability of Measure Properties

NQF's criteria for endorsement currently allow for testing reliability and validity for the data elements or the performance measure score, or ideally for both. The TEP recommended that composite performance measures should be tested at the level of the composite measure score as discussed below. Some TEP members also suggested that testing at the score level should be required for all individual performance measures as well as composite performance measures.

Reliability

Reliability is related to the probability of misclassification, and is therefore a key issue in performance measurement. 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 Individual performance measures that reflect the quality of care provided to patients by providers or institutions, when combined into a composite performance measure, should be useful in detecting a consistent pattern of practice or quality of care across patients of the provider or institution. That is, a composite performance measure is a set of measures that, taken together, are thought to reflect the quality of care, show a more consistent pattern within a provider's practice or within an institution than individual measures alone, and reveal 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 (i.e., differences in quality) from noise (i.e., random measurement error). Examples of analyses include signal-to-noise analysis,¹² interunit reliability,¹³ and intraclass correlation coefficient (ICC).¹⁴ Note that combining multiple indicators into a composite with all-or-none aggregation rules may not increase reliability of the performance measure score because the multiple indicators are essentially reduced to one data point on each patient.^e Therefore, all-or-none composite performance measures should demonstrate reliability of the composite measure score just as is recommended for any composite performance measure.

It is not essential that the individual component measure scores are reliable as long as the composite score itself is reliable. 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 is directed toward the inferences that can be made about accountable entities on the basis of their performance measure scores. For purposes of endorsing composite performance measures, validity testing of the constructed composite performance measure score is more important

^e Reliability of the performance measure is different than the rationale that all-or-none measures are intended to foster a system perspective of care, sometimes called "system reliability."

than validity testing of the component measures. 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 a true reflection of quality. However, TEP members noted that requiring empirical validity testing of the composite as a whole could be difficult to accomplish prior to NQF endorsement. Hence, if validity of the composite performance measure is not empirically demonstrated, then acceptable alternatives at the time of initial endorsement include: systematic assessment of content or face validity of the composite performance measure, or demonstration that each of the individual component measures meet the NQF subcriterion for validity. Empirical validity testing of the overall composite measure would be expected by the time of endorsement maintenance.

It may be unlikely that another valid measure (“gold standard”) 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 of composite performance measures is likely to focus on testing various theoretical relationships. For example, a composite measure that includes multiple process measures could be tested for its association with a measure of the outcome that those processes are intended to improve. Alternatively, a composite measure might be tested for its ability to predict future outcomes. Another approach is to test the ability of the composite performance measure to differentiate performance between groups known to differ on a similar or related quality construct.

Additional Testing to Support the Construction 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). Although this is listed as a separate criterion to signify that it is specific to just composite measures, it is in reality an extension of the reliability and validity subcriteria. For example, aggregation and weighting rules are intended to maximize reliability and discrimination. Missing data rules are intended to minimize bias (which means maximizing validity). The item on missing data is probably relevant to all performance measures, but is included in this criterion because the scope of this project was limited to composite measures, and because missing data problems may be magnified when multiple measures are aggregated.

This criterion 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 original wording of the criteria for testing composites was more relevant to composite measures that are based on correlated components. The modified criterion is intended to be neutral in terms of the analyses required. For example, if the quality construct and rationale for summarizing the component measures in a composite are based on their correlation with each other, then analyses based on shared variance (e.g., factor analysis, Cronbach’s alpha, item-total correlation, and mean inter-item correlation) are appropriate. In such cases, very high correlations between component measures may suggest that a component is redundant and not necessary, and very low correlations may indicate that the component measures do not reflect a common underlying quality construct. Conversely, if the quality construct and rationale for summarizing the measures in a composite are not based on their correlation with each other, then analyses demonstrating the contribution of each component to the composite (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable [15](#)), or their clinical justification (e.g., correlation of the individual component measures to a common outcome measure) are indicated. The TEP acknowledged that empirical analyses for composites with uncorrelated

component measures are not as well established as those for composites with correlated component measures. However, the TEP agreed that NQF should ask for the analyses, but allow for some flexibility if developers can make a strong case as to why empirical analyses were not conducted.

The unit of analysis for which performance measures are calculated is typically the service provider organization (hospital, clinic, health plan, etc.) rather than the individual patient. For such performance 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 support the construction 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.¹⁶⁻¹⁸

Although the primary purpose of this subcriterion is to justify the composite construction, parsimony and simplicity are noted as related objectives. Parsimony in regards to the number of component measures that are included and simplicity in regards to the aggregation and weighting rules help minimize burden (data collection, confusion, etc.). However, these related objectives are less important and should not compromise the conceptual integrity or reliability and validity of the composite measure.

Scientific Acceptability of Measure Properties is a must-pass criterion and measures must meet both reliability and validity. In addition, composite measures must meet the additional subcriterion for the composite performance measure in order to meet the must-pass criterion of Scientific Acceptability.

Feasibility

The standard feasibility criterion applies to the composite measure as a whole, but must take into account all of 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 criterion 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 with the use of the composite measure. Optimally, for purposes of improvement, the data should be collected and subsequently available to stakeholders at a granular level 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	
<p>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. <i>Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.</i></p> <p>1a. Evidence to Support the Measure Focus The measure focus is evidence-based, demonstrated as follows:</p> <ul style="list-style-type: none"> • 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. <p>AND</p> <p>1b. Performance Gap Demonstration of quality problems and opportunity for improvement, i.e., data ⁷ demonstrating</p> <ul style="list-style-type: none"> • considerable variation, or overall less-than-optimal 	<p>The evidence subcriterion (1a) must be met for each component of the composite (unless NQF-endorsed under the current evidence requirements). The evidence could be for a group of interventions included in a composite performance measure (e.g., studies in which multiple interventions are delivered to all subjects and the effect on the outcomes is attributed to the group of interventions).</p> <p>The performance gap criterion (1b) must be met for the composite performance measure as a whole. The performance gap for each component also should be demonstrated. However, if a component measure has little opportunity for improvement, justification for why it should be</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>performance, in the quality of care across providers; and/or</p> <ul style="list-style-type: none"> disparities in care across population groups. <p>AND</p> <p>1c. High Priority The performance measure addresses:</p> <ul style="list-style-type: none"> a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; <p>OR</p> <ul style="list-style-type: none"> 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). <p>Composite 1d. For composite performance measures, the following must be explicitly articulated and logical:</p> <ol style="list-style-type: none"> The quality construct, including the overall area of quality; included component measures; and the relationship of the component measures to the overall 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 the stated quality construct and rationale. 	<p>included in the composite is required (e.g., increase reliability of the composite, clinical evidence).</p> <p>The priority criterion (1c) applies to the composite performance measure as a whole.</p> <p>Subcriterion 1d must also be met for a composite performance measure to meet the must-pass criterion of Importance to Measure and Report.</p>
<p>2. Reliability and Validity—Scientific Acceptability of Measure Properties: Extent to which the measure, <u>as specified</u>, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. <i>Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.</i></p> <p>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</p>	<p>Add to Note 8: Composite measure specifications include component measure specifications (unless individually endorsed); scoring rules (i.e., how the component scores are combined or aggregated); how missing data are handled (if applicable); required sample sizes (if applicable); and when appropriate, methods for standardizing scales across component scores and weighting rules (i.e., whether all component scores are given equal</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>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.</p> <p>2b. Validity</p> <p>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.</p> <p>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.</p> <p>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</p>	<p>or differential weighting when combined into the composite).</p> <p>2a2. For composite performance measures, reliability must be demonstrated for the composite measure score. Testing should demonstrate that measurement error is acceptable relative to the quality signal. Examples of testing include signal-to-noise analysis ¹², interunit reliability¹³, and intraclass correlation coefficient.</p> <p>Demonstration of the reliability of the individual component measures is not sufficient. In some cases, component measures that are not independently reliable can contribute to reliability of the composite measure.</p> <p>2b2. For composite performance measures, validity should be empirically demonstrated for the composite measure score. If empirical testing is not feasible at the time of initial endorsement, acceptable alternatives include systematic assessment of content or face validity of the composite performance measure or demonstration that each of the component measures meet NQF subcriteria for validity. By the time of endorsement maintenance, validity of the composite performance measure must be empirically demonstrated. It is unlikely that a “gold standard” criterion exists, so validity testing generally will focus on construct validation – testing hypotheses based on the theory of the construct. Examples include testing the correlation with measures hypothesized to be related or not related; testing the difference in scores between groups known to differ on quality assessed by some other measure.</p> <p>2b3. Applies to the component measures and composite performance measures.</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>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</p> <p>2b4. For outcome measures and other measures when indicated (e.g., resource use):</p> <ul style="list-style-type: none"> • 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 <p>OR</p> <ul style="list-style-type: none"> • rationale/data support no risk adjustment/ stratification. <p>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.</p> <p>2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.</p> <p>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.</p> <p>Composite 2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:</p> <ol style="list-style-type: none"> 1) the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and 2) the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible; and 	<p>2b4. Applies to outcome component measures (unless NQF-endorsed).</p> <p>2b5. Applies to composite performance measures.</p> <p>2b6. Applies to component measures.</p> <p>2c. Applies to composite performance measures.</p> <p>Subcriterion 2d must also be met for a composite performance measure to meet the must-pass criterion of Scientific Acceptability of Measure Properties.</p> <p>If empirical analyses do not provide adequate results (or are not conducted), other justification must be provided and accepted for the measure to potentially meet the must-pass</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>3) the extent of missing data and how the specified handling of missing data minimizes bias (i.e., achieves scores that are an accurate reflection of quality).</p>	<p>criterion of Scientific Acceptability of Measure Properties.</p> <p>Examples of analyses:</p> <p>1) <i>If components are correlated</i> - analyses based on shared variance (e.g., factor analysis, Cronbach’s alpha, item-total correlation, mean inter-item correlation).</p> <p>1) <i>If components are not correlated</i> - analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable ¹⁵, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure).</p> <p>2) Ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.</p> <p>3) Overall frequency of missing data and distribution across providers. Ideally, sensitivity analysis of the effect of various rules for handling missing data and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.</p>
<p>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.</p> <p>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).</p>	<p>3a, 3b, 3c. Apply to composite performance measures as a whole, taking into account all component measures.</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>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.</p> <p>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).</p>	
<p>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.</p> <p>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</p> <p>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</p> <p>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).</p>	<p>Note that NQF endorsement applies only to the composite performance measure as a whole, not to the individual component measures (unless they are submitted and evaluated for individual endorsement).</p> <p>4a. Applies to composite performance measures. To facilitate transparency, at a minimum, the individual component measures of the composite must be listed with use of the composite measure.</p> <p>4b. Applies to composite performance measures.</p> <p>4c. Applies to composite performance measures and component measures. If there is evidence of unintended negative consequences for any of the components, the developer should explain how that is handled or justify why that component should remain in the composite.</p>

Measure Evaluation Criteria	Guidance for Composite Performance Measures
<p>5. Comparison to Related or Competing Measures If a measure meets the above criteria <u>and</u> 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.</p> <p>5a. The measure specifications are harmonized ²³ with related measures; OR the differences in specifications are justified.</p> <p>5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure); OR multiple measures are justified.</p>	<p>5a and 5b. Apply to composite performance measures as a whole as well as the component measures.</p>

Table 2. Notes to Measure Evaluation Criteria

Conditions
<p>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.</p> <p>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.</p>
1. Evidence, Performance Gap, and Priority—Importance to Measure and Report
<p>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.</p> <p>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).</p> <p>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.</p>

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 (if applicable), required sample sizes (if applicable); 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 measure 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 <i>disclosed and available</i> 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). <i>At a minimum, the data on performance results about identifiable, accountable entities are available to the public (e.g., unformatted database).</i> 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., <i>influenza immunization</i> of patients in hospitals or nursing homes); related measures with the same target population (e.g., eye exam and HbA1c for <i>patients with diabetes</i>); 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.

- NQF steering committees should include at least one member who is knowledgeable about composite performance measures and serve as the primary reviewer(s) of the composite performance measure 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.)

Next Steps

After these recommendations are approved by the Consensus Standards Approval Committee and ratified by the Board of Directors, several activities are required for implementation.

- The NQF measure submission form will be modified to identify composite performance measures and request the additional information required to evaluate composite performance measures based on the criteria and guidance in this report.
- The NQF measure evaluation criteria will be updated to incorporate the composite evaluation criteria and posted on the NQF web site.
- This guidance document will be posted to the NQF web page for submitting standards.
- NQF will present the approved guidance to measure developers.
- If composite performance measures are being considered for endorsement, calls for nominations to NQF steering committees will seek members with expertise in composite measure methods.
- NQF staff will be oriented to identifying composite performance measures to ensure that developers submit the information needed to evaluate composite performance measures.
- NQF staff will work to compile examples of composite measure submissions, including testing and analyses.

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. Booyesen 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.
10. Shwartz M, Ash AS. Composite measures: Matching the method to the purpose. *AHRQ 11-8-2008*; Available at: <http://www.qualitymeasures.ahrq.gov/expert/expert-commentary.aspx?id=16464>. Last accessed: March, 2013.
11. Edwards JR, Bagozzi RP, On the nature and direction of relationships between constructs and measures, *Psychol Methods*, 2000;5(2):155-174.
12. Adams JL, *The Reliability of Provider Profiling: A Tutorial*, Santa Monica, CA: RAND Corporation; 2009. Available at www.rand.org/pubs/technical_reports/TR653. Last accessed: January, 2011.
13. Zaslavsky AM, Statistical issues in reporting quality data: Small samples and casemix variation, *Int J Qual Health Care*, 2001;13(6):481-488.
14. Streiner DL, Norman GR, *Health Measurement Scales: A Practical Guide to Their Development and Use*, 4 ed., New York: Oxford University Press; 2008.

15. Diamantopoulos A, Winklhofer HM, Index construction with formative indicators: An alternative to scale development, *Journal of Marketing Research*, 2001;38(2):269-277.
16. O'Malley AJ, Zaslavsky AM, Hays RD, et al., Exploratory factor analyses of the CAHPS-Hospital pilot survey responses across and within medical, surgical, and obstetric services, *Health Services Research*, 2005;40(6p2):2078-2095.
17. O'Malley AJ, Zaslavsky AM, Domain-level covariance analysis for multilevel survey data with structured nonresponse, *Journal of the American Statistical Association*, 2008;103(484):1405-1418.
18. Zaslavsky AM, Cleary PD, Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey, *Med Care*, 2002;40(10):951-964.

Appendix A: Glossary

Term	Definition	Source
<p>All-or-None Scoring</p> <p><i>Also known as:</i></p> <ul style="list-style-type: none"> • <i>Appropriateness model</i> • <i>Conjunctive scoring</i> 	<p>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/or outcomes, and the numerator is the number of patients who actually received all of the care and/or outcomes for which the specific patient was eligible. No partial credit is given.</p>	<p>NQF Composite Guidance Report, 2009</p>
<p>Any-or-None Scoring</p>	<p>A percentage is determined by applying an any-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/or outcome, and the numerator is the number of patients who actually received any of the care or outcomes for which the specific patient was eligible. No partial credit is given.</p>	
<p>Bundle</p>	<p>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.ihl.org/IHI/Topics/CriticalCare/IntensiveCare/ImprovementStories/BundleUpforSafety.htm.</p>	<p>NQF Composite Guidance Report, 2009</p>
<p>Clinimetric Model</p>	<p>See "Formative Model".</p>	
<p>Component</p>	<p>A constituent part or element of a composite measure.</p>	<p>NQF Composite Guidance Report, 2009</p>
<p>Composite measure</p>	<p>A combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.</p>	<p>NQF Composite Guidance Report, 2013</p>
<p>Construct</p>	<p>An abstract phenomenon that is measured indirectly through less abstract indicators.</p>	<p>NQF Composite Guidance Report, 2009</p>

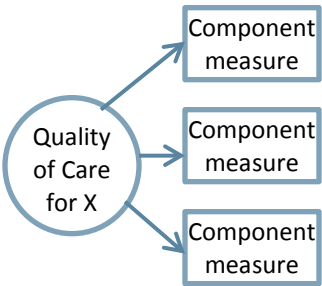
Term	Definition	Source
Domain	A dimension or aspect of a construct.	NQF Composite Guidance Report, 2009
Formative Model	<p>A conceptual model in which a set of indicators are combined to form or cause the construct of interest; that is, the individual indicators form or define the quality construct.</p> <p>For example, stressful life events such as going to jail, buying a house, having a spouse die, all cause stress; they aren't (hopefully) related to each other for any individual, but may cause more stress.</p>	Edwards & Bagozzi, 2000
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, 2009
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, 2009
Latent variable	An unobserved trait or characteristic	NQF Composite Guidance Report, 2009
Measure	Numeric quantification of some concept. A quality measure is a numeric quantification of healthcare quality.	NQF Composite Guidance Report, 2009

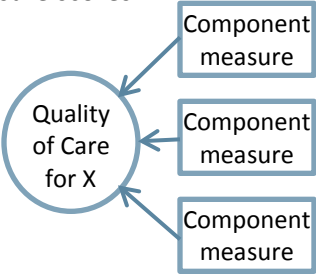
Term	Definition	Source
Opportunity scoring	<p>Scoring used with process measures, determined from the sum of all numerators (achieved the desired process) divided by the sum of all denominators (i.e., number of eligible patients or opportunities, which could vary by measure).</p> <p>If the opportunity 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)].</p> <p>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”.)</p>	NQF, Composite Guidance Report, 2009, 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, 2009
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 model	See "Reflective Model".	

Term	Definition	Source
Quality construct	<p>A concept of quality. For a composite performance measure it includes:</p> <ul style="list-style-type: none"> • the overall area of quality (e.g., quality of CABG surgery); • the included component measures (e.g., pre-operative beta blockade; CABG using internal mammary artery; CABG risk-adjusted operative mortality); • the conceptual relationships between each component and the overall composite (e.g., components cause or define quality, components caused by or reflect quality); and • the relationships among the component measures (e.g., correlated or not, process leads to outcome). 	NQF Composite Guidance Report, 2013
Reflective Model	<p>A conceptual model in which the quality construct can be viewed as the cause of individual indicators; that is, the individual indicators reflect or manifest the quality construct.</p> <p>For example, feeling nervous or anxious, feeling overwhelmed, feeling unable to cope, etc. can all be caused by stress and more of those feelings indicate higher levels of stress.</p>	Edwards & Bagozzi, 2000
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, 2009
Subscale	A measure of a dimension of a scale composed of a subset of the items in a scale.	NQF Composite Guidance Report, 2009
Variable	A characteristic or attribute that varies within and among people or the subjects of study.	NQF Composite Guidance Report, 2009

Appendix B: Approaches for Constructing Composite Performance Measures

A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score. Following are descriptions of some of the more common approaches for constructing composite performance measures. The examples for analyses are preliminary and will be expanded in the future.

Quality Construct	Description
<p>1. The quality construct is seen as causing or reflected in the component measure scores</p>  <pre> graph LR A((Quality of Care for X)) --> B[Component measure] A --> C[Component measure] A --> D[Component measure] </pre> <ul style="list-style-type: none"> • Also known as reflective, psychometric, scale, homogenous scale, dimensional • Example: NQF#0696: CABG Composite Score (STS) 	<ul style="list-style-type: none"> • Scores on the component measures are considered the effect of quality (or caused by quality) • Component 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 measures should be correlated with one another because they share common variance due to relationship with the construct; and each component is correlated with the total composite score (omitting the component being assessed) • Analyses based on shared variance(e.g., factor analysis, Cronbach’s alpha, item-total correlation, and mean inter-item correlation) support the construction of the composite. <p>Aggregation Examples:</p> <p>Combination of multiple individual performance measures Various approaches may be used, including:</p> <ul style="list-style-type: none"> ▪ 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.)

Quality Construct	Description
<p>2. The quality construct is seen as being caused or defined by the component measure scores</p>  <pre> graph LR CM1[Component measure] --> QCC((Quality of Care for X)) CM2[Component measure] --> QCC CM3[Component measure] --> QCC </pre> <ul style="list-style-type: none"> • Also known as formative, clinimetric, index, heterogeneous index, categorical <p>Example: NQF# 0530: Mortality for Selected Conditions (AHRQ)</p>	<ul style="list-style-type: none"> • Component measures are considered to cause (or define) quality • Component measures define the quality construct and should cover the scope of the quality construct; therefore, focusing QI on the component performance measures should change the composite score • Component measures do not need to be correlated with one another but could be correlated (correlation between components could be zero, positive, or negative) • Analyses based on shared variance are not consistent with this model. Analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable ¹⁵, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure) support the construction of the composite. <p>Aggregation Examples:</p> <p>Combination of multiple individual performance measures Various approaches may be used, including:</p> <ul style="list-style-type: none"> ▪ 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.)

Quality Construct	Description
<p>3. The quality construct is viewed or defined as receiving all necessary care represented by the component measures</p> <p>3a. All components must be achieved to signal quality. Failure on any component is viewed as a failure.</p> <ul style="list-style-type: none"> • Also known as All-or-None <p>Example: NQF# 0729: Optimal Diabetes Care (MN Community Measurement)</p> <p>3b. The more components achieved, the greater the quality signal</p> <ul style="list-style-type: none"> • Also known as partial credit, percentage of necessary care 	<ul style="list-style-type: none"> • Component measures define the quality construct and should cover the scope of the quality construct. • Component measures represent multiple care processes (foot care, eye care, glucose control), not linked steps in one care process (assess immunization status, counsel patient, and administer vaccination). • Component measures are assessed for each patient. • Analyses demonstrating the contribution of each component to the composite score (e.g., frequency of failure on each component); or correlation of the individual component measures to a common outcome measure support the construction of the composite. <p>Aggregation Examples:</p> <p>3a. 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]</p> <p>3b. 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]</p>

Quality Construct	Description
<p>4. The quality construct is viewed as individual patients not experiencing any healthcare-acquired adverse event/complication or not receiving unnecessary or inappropriate care.</p> <ul style="list-style-type: none"> • Also known as any-or-none <p>Example: NQF# 0564: Complications within 30 Days Following Cataract Surgery Requiring Additional Surgical Procedures (PCPI)</p>	<ul style="list-style-type: none"> • Component measures define the quality construct and should cover the scope of the quality construct. • Component measures are assessed for each patient. • Analyses demonstrating the contribution of each component to the composite score (e.g., frequency of occurrence on each component); or correlation of the individual component measures to a common outcome measure support the construction of the composite. <p>Aggregation Examples:</p> <p>Composite numerator - Multiple components specified in the numerator and measured for each patient</p> <p>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]</p>

Appendix C: References Consulted

1. Adams JL, *The Reliability of Provider Profiling: A Tutorial*, Santa Monica, CA: RAND Corporation; 2009. Available at www.rand.org/pubs/technical_reports/TR653. Last accessed: January, 2011.
2. 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.
3. Asch S, Hofer T, Representing overall quality of care: The whole must be more than the sum of the parts, White Paper: Advancing Quality Measurement Conference, 2008. (Unpublished)
4. 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.
5. Booyesen F, An overview and evaluation of composite indices of development, *Social Indicators Research*, 2002;59:115-151.
6. Diamantopoulos A, Winklhofer HM, Index construction with formative indicators: An alternative to scale development, *Journal of Marketing Research*, 2001;38(2):269-277.
7. Dijkers MP, Psychometrics and clinimetrics in assessing environments. A comment suggested by Mackenzie et al., 2002, *J Allied Health*, 2003;32(1):38-43.
8. 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.
9. 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.
10. 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.
11. Edwards JR, Bagozzi RP, On the nature and direction of relationships between constructs and measures, *Psychol Methods*, 2000;5(2):155-174.
12. 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.
13. 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.
14. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, et al., Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns, *Journal of Clinical Epidemiology*, 2007;60;pp. 651-657.

15. 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.
16. 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.
17. Ingenix, Creating quality composite scores: Challenges and issues in physician quality measurement, 2008; October 2012. Available at http://www.optuminsight.com/content/attachments/29504_Decision_Support_EBM_WhitePaper_LO4.pdf . Last accessed: March, 2013.
18. Institute of Medicine, *Performance Measurement: Accelerating Improvement*, Washington, DC: National Academies Press; 2006.
19. Jacobs R, Goddard M, Smith PC, How robust are hospital ranks based on composite performance measures?, *Med Care*, 2005;43(12):1177-1184.
20. 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.
21. Kaplan SH, Normand SL, *Conceptual and Analytical Issues in Creating Composite Measures of Ambulatory Care Performance*, Washington, DC: National Quality Forum; 2006.
22. 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.
23. Kianifard F, Evaluation of clinimetric scales: Basic principles and methods, *The Statistician*, 1994;43(4):475-482.
24. 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.
25. National Committee for Quality Assurance (NCQA), MEMO: Summary of Alliance Use of Composite Measures, 2010. Available at <http://www.rwjf.org/content/dam/web-assets/2010/06/summary-of-alliance-use-of-composite-measures> . Last assessed: March, 2013.
26. 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.
27. Nolan T, Berwick DM, All-or-none measurement raises the bar on performance, *JAMA*, 2006;295(10):1168-1170.
28. Normand C, Measuring outcomes in palliative care: limitations of QALYs and the road to PalYs, *J Pain Symptom Manage*, 2009;38(1):27-31.

29. 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.
30. O'Malley AJ, Zaslavsky AM, Hays RD, et al., Exploratory factor analyses of the CAHPS-Hospital pilot survey responses across and within medical, surgical, and obstetric services, *Health services research*, 2005;40(6p2):2078-2095.
31. O'Malley AJ, Zaslavsky AM, Domain-level covariance analysis for multilevel survey data with structured nonresponse, *Journal of the American Statistical Association*, 2008;103(484):1405-1418.
32. 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.
33. 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.
34. 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.
35. 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.
36. Ross JS, Correlation of inpatient and outpatient measures of stroke care quality within veterans health administration hospitals, 2011;42(8):2269-2275.
37. 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.
38. 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.
39. 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.
40. Shwartz M, Ash AS. Composite measures: Matching the method to the purpose. *AHRQ 11-8-2008*; Available at: <http://www.qualitymeasures.ahrq.gov/expert/expert-commentary.aspx?id=16464>. Last accessed: March, 2013.
41. Staiger DO, Dimick JB, Baser O, et al., Empirically derived composite measures of surgical performance, *Med Care*, 2009;47(2):226-233.

42. Streiner DL, Clinimetrics vs. psychometrics: an unnecessary distinction, *J Clin Epidemiol*, 2003;56(12):1142-1145.
43. Streiner DL, Being inconsistent about consistency: when coefficient alpha does and doesn't matter, *J Pers Assess*, 2003;80(3):217-222.
44. Streiner DL, Norman GR, *Health Measurement Scales: A Practical Guide to Their Development and Use*, 4 ed., New York: Oxford University Press; 2008.
45. 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.
46. 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.
47. Zaslavsky AM, Statistical issues in reporting quality data: Small samples and casemix variation, *Int J Qual Health Care*, 2001;13(6):481-488.
48. Zaslavsky AM, Cleary PD, Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey, *Medical Care*, 2002;40(10):951-964.

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

National Quality Forum
1030 15th St NW, Suite 800
Washington, DC 20005
<http://www.qualityforum.org>

Contract HHSM-500-2009-00010C
Task order 8

ISBN 978-1-933875-52-1
©2013 National Quality Forum