NATIONAL QUALITY FORUM

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

TO: Composite Expert Panel

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

SU: Briefing materials for November 2, 2012 in-person meeting

DA: October 29, 2012

We look forward to seeing you on November 2 for the in-person meeting of the Expert Panel.

The purpose of the meeting is to:

- Identify appropriate evaluation methods for various types of composite performance measures.
- Identify any unique considerations for evaluating composite performance measures in relation to NQF's endorsement criteria.
- Develop guidance for evaluating and submitting composite performance measures for NQF endorsement.

CONTENTS

CONTEXT

The overall purpose of the Composite Measure Evaluation Guidance project is to review and update NQF's guidance on evaluating composite performance measures for potential NQF endorsement.

Although composite performance measures may be developed with a particular purpose in mind, all NQF-endorsed measures are considered suitable both for performance improvement and accountability. NQF does not currently endorse performance measures for specific accountability applications (e.g., payment vs. public reporting).

The term "composite measure" may refer to scales or instruments to assess individuals (e.g., PHQ-9, CAHPS) or to performance measures used to assess healthcare providers. However, NQF endorses only performance measures; it does not endorse the instruments and scales used with individuals. NQF's recent project Patient-Reported Outcomes (PROs) in Performance Measurement addresses performance measures based on PROs (health-related quality of life including functional status, symptoms and symptom burden, experience with care, and health-related behaviors).

NQF EXPERIENCE WITH COMPOSITE PERFORMANCE MEASURES

After our initial composite project, composite measures were submitted and evaluated in projects by topic area. The composite information was the last form added to NQF's online measure submission form and the supplemental composite form was not submitted for all composite measures that are currently endorsed. Following are some preliminary statistics on NQF's experience with composite performance measures. We will provide more detail at the in-person meeting.

A total of 27 composite measures have been submitted to NQF. Of these, 21 are currently endorsed. Of the 21 endorsed measures, two are all-or-none measures, six are based on CAHPS surveys, another four are based on other patient/family surveys, and the remaining 15 include various combinations of individual performance measures (with varying rationales, aggregation methods, weighting schemes, etc.).

The reasons that six composite performance measures were not endorsed varied by measure and included:

- lack of variability and overall high performance on the composite performance score;
- lack of evidence supporting the components of an all-or-none measure;
- withdrawn from consideration by the measure steward;
- component performance measures were not endorsed and did not meet criteria;
- composite measures included some component performance measures that lost endorsement and/or missing data had a substantial impact; and
- composite measures did not include component measures that were more representative of quality of care.

APPROPRIATE METHODS OF EVALUATING VARIOUS TYPES OF COMPOSITE PERFORMANCE MEASURES

NQF's definition of a composite performance measure is:

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

The term "composite measure" has been applied to many types of measures; however, NQF's current guidance for their evaluation does not adequately address the various types (e.g., all-or-none vs. multiple individual performance measures).

For reviewing appropriate methods for evaluating composite performance measures, we have identified three categories of measures:

- composite performance measures that may require additional evaluation and criteria beyond what is required for all performance measures (i.e., those composed of multiple individual performance measures; see Table 1);
- composite performance measures that may not require additional evaluation and criteria beyond what is required for all performance measures (i.e., all-or-none, statistical models with shrinkage estimators; see Table 2); and
- performance measures that are not composites or composite measures that are not performance measures and thus not considered for NQF endorsement (i.e., used at individual level of assessment; see Table 3).

For this stage of the TEP's review and discussion, the focus will be considerations for measure evaluation rather than selecting specific terms or taxonomy. After all issues have been discussed, the TEP may consider whether these categorizations as presented are useful for a taxonomy, or whether modifications should be made to the categorizations (for example, Alan's chart is included at the end of the appendix). However, this is not the primary objective of the project.

Composite Performance Measures and Conceptual Models

Much of the literature on composite measures is focused on the development of multi-item scales or indexes used with individuals (e.g., depression scale, functional status scale, Apgar score). The principles of composite measurement are applied to performance measure composites, where the unit of analysis is a healthcare provider rather than an individual, and the components are aggregate scores on performance rather than responses to questions.

The literature, and the TEP on the initial conference call, indicates the importance of the conceptual model for composite measure development and eventual validation. The literature refers to two major conceptual approaches to composite measures. They are given different names depending on the discipline. Both models are measuring an abstract concept, in this case quality of care. The basic differences lie in the relationship among the components and between the components and the abstract quality construct (or composite score).

- Model 1 is the traditional conceptual model where the abstract construct causes the observed scores on the component indicators. Thus a person with severe depression will respond to items in a different way than someone without depression. A potential translation to quality performance might be that an organization with a strong safety culture will have higher scores on individual safety performance measures than an organization without a safety culture. In this model, the components have to be correlated because they have a shared common variance and traditional psychometric analyses of factor analysis, inter-item correlation, and Cronbach's alpha for internal consistency reliability apply. The components are considered a random sample of possible indicators and therefore do not have to be all-inclusive.
- In Model 2, the observed scores on the component indicators are thought to cause or define the underlying abstract construct. Thus, an Apgar score that measures the state of health of a newborn is based on various items. A potential translation to quality performance might be a set of process performance measures that define quality of care (e.g., aspirin, fibrinolytic, and PCI within 90 minutes to define quality care for acute AMI). In this model, the components do not need to be correlated. The difficulty with this model is that the components need to cover the entire scope of quality and omitted components impact validity. The numerator components for all-or-none composites also are intended to define quality, but are focused on those that are absolutely necessary and the measure is constructed differently.
- Some composite performance measures are not based on a conceptual model, and instead use what is available regardless of the conceptual fit and the only requirement is that it is a good empirical predictor of some measurable gold standard or future state.

Key Questions for Guided General Discussion

- Does the conceptual model for a composite performance measure dictate measure construction and methods of evaluation, and if so, how?
- Selection of component performance measures influence reliability and validity of the composite performance measure. However, if reliability and validity of the final composite performance measure score is adequately demonstrated, are additional analyses of the components and structure of the composite necessary?
- What are the primary issues regarding selection of component performance measures and their conceptual relationships that need to be addressed?
- Can analyses such as factor analysis and internal consistency reliability be applied when the unit of analysis is providers (vs. people) and the data are performance measure scores (vs. item responses)?
- If a composite performance measure does not fit either of the main conceptual approaches delineated in Table 1, what is appropriate justification? (e.g., composite just includes what's available and the component performance measures are not correlated and also do not cover the scope of the quality construct)
- What is the conceptual model for composite performance measures with a main purpose to increase reliability? (Does the principle of increased reliability with increased number of items hold for all-or-none measures when components are reduced to one data point?)

Composite Performance Measures that May Require Additional Analyses/Evaluation Criteria

Table 1 includes composite performance measures composed of multiple individual level performance measures. These are the types of composite performance measures for which the current composite evaluation guidance is most relevant. Two conceptual models are described and unique considerations for evaluation are proposed. Specific questions are posed for the TEP to address when developing guidance on evaluating composite performance measures. ***Unless specifically mentioned, all basic NQF measure evaluation criteria apply to the composite performance measure.***

DRAFT Table 1. Composite Performance Measures that May Require Additional Analyses/Evaluation Criteria

Conceptual Model	Description	Unique Considerations for Evaluation
1. Conceptual model considers the quality construct as causing the effect on the observed component performance measure scores • Also known as psychometric, reflective, scale, homogenous scale • Example: NQF# 0530: Mortality for Selected Conditions (AHRQ)	• **Combination of multiple individual performance measures** • Component performance measures are considered effect indicators or variables • Component performance measures are considered a random sample of indicators of quality and should be interchangeable • Therefore, focusing QI only on the component performance measures may not change the composite score • Component performance measures should be correlated because they share common variance	• Evidence–Conceptual basis for component performance measures based on evidence of relationship with desired outcomes – i.e., NQF basic evidence criterion applies to component performance measures (evidence evaluation waived if component performance measure is already NQF- endorsed) • Reliability and validity testing– NQF basic criteria apply except that testing should be conducted at the level of the composite performance

Conceptual Model	Description	Unique Considerations for Evaluation
	Structure: ***Does conceptual model dictate structure of the composite measure?*** ***Can any of these methods be used with this conceptual model?*** ▪ 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.)	score (not at the data element level) ***What are examples of validity testing – correlation with desired outcomes or gold standard measure if exists; difference between known-groups?*** • Additional evaluation: evaluation of the composite structure and components follows traditional psychometric approaches ***Can these methods be applied when the unit of analysis is providers and the data are performance measure scores?*** ▪ Inter-item correlation between component performance measures ▪ Factor analysis to indicate unidimensionality ▪ Internal consistency reliability ***What other additional evaluation is indicated for the measure construction– e.g., weighting scheme whether equal or differential; other?***
2. Conceptual model considers the observed performance measure scores as causing (or defining) the quality construct • Also known as clinimetric, formative, index, heterogenous index • Example: NQF#0696: STS CABG Composite	• **Combination of multiple individual performance measures** • Component performance measures are considered causal indicators or variables • Component performance measures define the quality construct and must cover the entire scope of the quality construct • Component performance measures do not need to be correlated **Structure:** ***Does conceptual model dictate structure of the composite measure?*** ***Can any of these methods be used with this conceptual model?*** ▪ Opportunities (sum of all numerators / sum of all denominators) ▪ Average/weighted average of component measure scores (score on	• Evidence–Conceptual basis for component performance measures based on evidence of relationship with desired outcomes – i.e., NQF basic evidence criterion applies to component performance measures (evidence evaluation waived if component performance measure is already NQF- endorsed) • Reliability and validity testing– NQF basic criteria apply except that testing should be conducted at the level of the composite performance score (not at the data element level) ***What are examples of validity testing – correlation with desired outcomes or gold standard measure if exists; difference between known-groups?***

Conceptual Model	Description	Unique Considerations for Evaluation
	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.)	• Additional evaluation: Component evaluation ***When traditional psychometric analyses are not indicated, what is appropriate?*** ▪ Each component performance measure should be correlated with some variable external to the composite (e.g., desired outcome) ▪ Each component performance measure meets criteria for reliability (or is already NQF endorsed) Is reliability of data elements sufficient for the component performance measures? ▪ Content validity to ensure the scope of the quality construct is covered ***What other additional evaluation is indicated for the measure construction– e.g., weighting scheme whether equal or differential; other?***
Other?		

Additional Questions:

- Could "balancing" measures be included in a composite? A balancing measure is not the main focus of interest but is used to identify adverse consequences. For example, a performance measure about treating substance use that requires identification of patients with substance use problems will not be accurate if most patients are not even screened. Therefore, the screening measure might be considered a balancing measure. As another example, when measuring outcomes, a measure of case mix severity could be a balancing measure for concern about adverse patient selection. Some balancing measures would meet NQF criteria. For example, readmission might be a balancing measure for a performance measure on average length of stay.
- Should process and outcome measures be included a composite, and if so which conceptual model is used? Because health outcomes are integrative, reflecting the influence of multiple care processes and disciplines involved in the care, a composite of the process measures *paired* with a separate health outcome measure may be more conceptually consistent with the evidence and also may be more interpretable. Intermediate clinical outcomes and health outcomes might be considered differently.

Composite Performance Measures that May Not Require Additional Analyses/Evaluation Criteria

Table 2 includes composite measures that are characterized with multiple numerator components that are measured for each patient. The additional composite measure evaluation criteria identified in Table 1 may not be appropriate for these types of composite performance measures. Often these types of measures have been submitted on an individual measure form. Three conceptual models are described and unique considerations for evaluation are proposed. Specific questions are posed for the TEP to address when developing guidance on evaluating composite performance measures. ***Unless specifically mentioned, all basic NQF measure evaluation criteria apply to the composite performance measure.***

DRAFT Table 2. Composite Performance Measures that May Not Require Additional Analyses/Evaluation Criteria

Conceptual Model	Description	Unique Considerations for Evaluation
3. Conceptual model considers the quality construct as receiving all necessary care • Also known as All-or-None • Example: NQF# 0729: Optimal Diabetes Care (MN Community Measurement)	• **Composite numerator** - Multiple components specified in the numerator and measured for each patient • Percentage of patients who received all necessary components of care **Structure:** # of patients in the denominator who met all components (A **and** B **and** C **and** . . .) / # of patients in target population	• Evidence –NQF basic evidence criterion requires empirical evidence that the measured process influences a desired outcome. This would apply to each of the numerator components. However, in some cases the evidence may be for the set of components together rather than for each component individually. • Reliability and validity testing– NQF basic criteria allow for testing at the level of data elements or performance measure score. ***Should testing be conducted at level of composite performance measure score?*** • ***Are any additional analyses of the component numerators needed?***
4. Conceptual model considers the quality construct as receiving necessary care, but receiving some is better than none • Also known as partial credit, percentage of necessary care • Example: possibly NQF# 0731: Comprehensive Diabetes Care (NCQA)?	• **Composite numerator** - Multiple components specified in the numerator and measured for each patient • Average percentage of necessary components of care received by patient **Structure**: Sum of percentage of components met (A, B, C . . .) for each patient in the denominator / # of patients in target population	• Same as above • ***Are there any differences in evaluation of models 1 and 2?*** • ***Are any additional analyses of the component numerators needed?***

Conceptual Model	Description	Unique Considerations for Evaluation
5. Conceptual model considers the quality construct as not experiencing any healthcare-acquired adverse event/complication • Also known as any-or-all Example: NQF# 0564: Complications within 30 Days Following Cataract Surgery Requiring Additional Surgical Procedures (PCPI)	• **Composite numerator** - Multiple components specified in the numerator and measured for each patient • Percentage of patients who experienced any of the component adverse events or complications **Structure:** # of patient in the denominator who experienced A **or** B **or** C **or** . . . . / # of patients in target population	• Same as above except NQF basic evidence criterion treats measures of health outcomes differently and requires only a rationale that it is related to at least one healthcare structure, process, intervention, or service. • ***Are there any differences in evaluation from models 1 and 2?*** • ***Are any additional analyses of the component numerators needed?***
6. Conceptual model considers the quality construct defined by one concept but uses additional information on average performance to increase precision (reliability) Also known as reliability adjustment, shrinkage estimator Example: NQF# 0737: Survival Predictor for Esophagectomy Surgery (Leapfrog) *Note: NQF did not consider this a composite performance measure at the time of review*	• Combines two rates of the same concept (e.g., a provider's observed mortality rate and an average mortality rate for a specific category of providers such as quartile by patient case volume) • To-date has been used only with outcome measures • Uses a provider characteristic to categorize all providers for purposes of creating an average rate (in some cases, case volume was an endorsed performance measure and used as the characteristic to categorize providers) **Structure:** (Weight x observed rate) + (weight x average rate) Weight is based on reliability of the provider observed rate, which is influenced by case volume	• Evidence –NQF basic evidence criterion applies • Reliability and validity testing– NQF basic criteria allow for testing at the level of data elements or performance measure score. ***Should testing be conducted at level of composite performance measure score?*** • NQF basic criteria for risk adjustment applies. • ***If the shrinkage target is identified for specific groups of providers, what justification must be provided for the selected characteristic?*** ▪ Association with outcome ▪ Not confounded with quality of care

Measures that are Not Composite Performance Measures

Table 3 includes examples of measures that are not considered composite performance measures, or in one case not even a performance measure that would be considered for NQF endorsement.

DRAFT Table 3. Measures that are Not a Composite *Performance* Measure

Conceptual Model	Description	Unique Considerations for evaluation
7. Conceptual models 1 and 2 traditionally used with multi-item composites to measure individuals Example: Model 1-PHQ-9, CAHPs; Model 2-Apgar	• Multi-item scale, instrument, index, survey administered to individuals. • Patient data on these scales may be used in an individual performance measure or a composite performance measure; but the scale itself is not a performance measure and not eligible for NQF endorsement.	• Not a composite *performance* measure • If patient data from such a scale is used in a performance measure, the reliability and validity of the scale also must be demonstrated. • See PRO project.
8. Multiple aspects of quality are identified, but there is no conceptual model that combines them to represent a quality construct Example: NQF# 0101 Falls: Screening, Risk-Assessment, and Plan of Care to Prevent Future Falls (NCQA)	There are two variants: • Separate measures that are identified to be reported together • Multiple related measures submitted on one form, but require computation of individual performance measure scores; some of the measures submitted on one form may have multiple denominators as well as numerators	• Some have done this when identification of a target population for an intervention requires something besides typical diagnosis (e.g., need to assess smoking status to identify those who need cessation counseling). • Typically, the assessment measure is less likely to meet the importance criteria. • Individual performance measures should be submitted on separate forms and evaluated individually against the measure evaluation criteria.

EVALUATION CRITERIA FOR COMPOSITE PERFORMANCE MEASURES

DRAFT Table 4. NQF Endorsement Criteria and Special Considerations Related to Composites

Abbreviated NQF Endorsement Criteria	Current Additional Criteria for Composites	Considerations
Importance to Measure and Report a. High impact b. Opportunity for improvement c. Health outcome OR evidence-based process/structure of care	The individual measures included in the composite or subcomposite measures must be either: NQF endorsed *OR* assessed to have met the individual measure evaluation criteria as the first step in evaluating the composite measure. (This does not apply to subscales of a scale/ instrument that cannot be used independently of the total scale.) Composite. 1d. The purpose/objective of the composite measure and the construct for quality are clearly described. Composite. 1e. The component items/ measures (e.g., types, focus) that are included in the composite are consistent with and representative of the conceptual construct for quality represented by the composite measure. Whether the composite measure development begins with a conceptual construct or a set of measures, the measures included must be conceptually coherent and consistent with the purpose.	Need to be more specific about the individual criteria Is this a criterion? Relevant to every performance measure – description and rationale This is difficult to apply and duplicative of 2i-2j
Scientific Acceptability of Measure Properties a. Reliability 1. precise specifications 2. reliability testing for either data elements or performance measure score b. Validity 1. specifications consistent with evidence 2. validity testing for either data elements or performance measure score 3. exclusions 4. risk adjustment 5. identify differences in	Composite specifications include methods for standardizing scales across component scores, scoring rules (i.e., how the component scores are combined or aggregated), weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite), handling of missing data, and required sample sizes. Composite. 2i. Component item/ measure analysis (e.g., various correlation analyses such as internal consistency reliability), demonstrates that the included component items/measures fit the conceptual construct; *OR* justification and results for alternative analyses are provided. Composite. 2j. Component item/ measure analysis demonstrates that the included components contribute to the variation	What are the required specifications? Replace 2i. -2k. with one criterion to be inserted in NQF basic criteria (similar to risk adjustment for outcome measures) For composite measures that combine the computed scores from individual performance measures: the conceptual model

Abbreviated NQF Endorsement Criteria	Current Additional Criteria for Composites	Considerations
performance 6. comparability of multiple data sources	in the overall composite score; *OR* if not, justification for inclusion is provided. Composite. 2k. The scoring/ aggregation and weighting rules are consistent with the conceptual construct. (Simple, equal weighting is often preferred unless differential weighting is justified. Composite. 2l. Analysis of missing component scores supports the specifications for scoring/ aggregation and handling of missing component scores.	is identified and analyses appropriate to the model justify the specified component performance measures.
Feasibility a. Data generated and used in care delivery b. Electronic data c. Data collection strategy can be implemented		
Usability and Use a. Accountability and transparency b. Improvement c. Benefits outweigh unintended negative consequences	Composite. 3d. Data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding. Composite. 3e. Demonstration (through pilot testing or operational data) that the composite measure achieves the stated purpose/objective.	This is an implementation issue – difficult to require unless require specifications to show individual component scores New usability and use subcriterion requires data on improvement – what else would be needed?

APPENDICES

Appendix A—Glossary

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

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