Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement

Effective September 2019

Contents

Changes from the 2018 Criteria and Guidance

This document updates the 2018 Measure Evaluation Criteria and Guidance.

- Updated terminology related to eCQM specifications
- Provided clarification regarding requirements for testing of measures based on ICD-10 coding
- Added guidance for reporting signal-to-noise reliability estimates
- Added guidance for describing the methods, results, and interpretation of construct validation
- Updated evaluation criteria for cost and resource use measures
- Updated the Scientific Methods Panel Frequently Asked Questions (FAQs)

Introduction

This document contains the measure evaluation criteria as well as additional guidance for evaluating measures based on the criteria. Additional information is available in detailed reports that can be accessed through NQF's Submitting Standards webpage.

Conditions for Consideration

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

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

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

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

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

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

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

Criteria for Evaluation

If all conditions for consideration are met, measures are evaluated for their suitability based on standardized criteria in the following order:

1. *Importance to Measure and Report*
2. *Scientific Acceptability of Measure Properties*
3. *Feasibility*
4. *Usability and Use*
5. *Related and Competing Measures*

Not all acceptable measures will be equally strong on each set of criteria. The assessment of each criterion is a matter of degree. However, if a measure is not judged to have met minimum requirements for *Importance to Measure and Report, Scientific Acceptability of Measure Properties*, and *Use* (first subcriterion under *Usability and Use*), it cannot be recommended for endorsement and will not be evaluated against the remaining criteria. These criteria apply to all performance measures (including outcome and resource use measures, instrument-

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

[2] An eCQM that has not been tested sufficiently to meet endorsement criteria may be eligible for Approval for Trial Use. Time-limited endorsement is no longer available.

based measures, composite performance measures, and eCQMs), except where indicated for a specific type of measure.

For **composite performance measures**, the following subcriteria apply to each of the component measures: 1a; 1b (also composite); 2b2 (also composite); 2b3; 2b5; 2b6; 4b2 (also composite); 5a and 5b (also composite).

Categories and Types of Measures

Healthcare performance measures are used to quantify healthcare processes, outcomes, patient (or other respondent) perceptions, and organizational structure and/or systems that are associated with the ability to provide high-quality care. There are four main "categories" of performance measures: Quality, Cost & Resource Use, Efficiency, and Access. Within these categories, there may be three main "types" of measures: Structure, Process, and Outcome. Within each of the measure types, there may be additional "sub-types" that further describe the measure. Another type of performance measure that combines two or more individual performance measures into a single measure with a single score; this type of measure is a "composite" measure. A composite measure may include any combination of measure types.

Quality measures assess performance on the six healthcare aims specified by the IOM: safety, timeliness, effectiveness, efficiency, equity, and patient centeredness.

Cost/resource measures are broadly applicable and comparable measures of health services counts (in terms of units or dollars) that are applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of use of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use.

Efficiency measures combine the concepts of resource use and quality. NQF has defined efficiency broadly as the resource use (or cost) associated with a specific level of performance with respect to the other five Institute of Medicine (IOM) aims of quality: safety, timeliness, effectiveness, equity, and patient-centeredness.

Access measures assess the ability to obtain needed healthcare services in a timely manner, including the perceptions and experiences of people regarding their ease of reaching health services or health facilities in terms of proximity, location, time, and ease of approach. Examples may include, but are not limited to, measures that address the timeliness of response or services, time until next available appointment, and availability of services within a community.

Structure of care is a feature of a healthcare organization or clinician related to its capacity to provide high-quality healthcare.

Process of care is a healthcare-related activity performed for, on behalf of, or by a patient. Appropriate Use is a type of process measure that has been used to evaluate procedures and medical technologies. Appropriate use measures are neither cost/resource use measures nor efficiency measures.

Outcome – An outcome of care is the health status of a patient (or change in health status) resulting from healthcare—desirable or adverse. A patient-reported outcome (PRO) is any report of the status of a patient's (or person's) health condition, health behavior, or experience with healthcare that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. Key PRO domains include health-

related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.

Intermediate clinical outcome is a change in physiologic state that leads to a longer-term health outcome.

Composite measures combine two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score. For the purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composite performance measures:

- 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, including all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient).

Instrument-based performance measures use data derived from instruments. "Instrument" is a generic term that researchers use for a measurement device (e.g. survey, test, questionnaire, scale). Instruments are used for consistently obtaining (or presenting) data from respondents. The data derived from an instrument may include ratings or ranking output that is included in the calculation of a performance measure. Instruments may be used to collect information from a variety of individuals; examples include patients, observers (e.g., family, or other caregivers), or clinicians. Data from instruments can be used in the calculation of structure, process, or outcome performance measures. Instruments specific to patient-reported outcomes may be referenced as PROMs (patient-reported outcome measures).

NQF Measure Evaluation Criteria[3]

1. Importance to Measure and Report: Evidence and Performance Gap
Extent to which the specific measure focus is evidence-based and important to making significant gains in healthcare quality where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria*.**

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

- Outcome: Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 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 that the measured structure leads to a desired health outcome.
- Efficiency: evidence is required for the quality component but not required for the resource use component. (Measures of efficiency combine the concepts of resource use and quality.)
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful (see Table 13 under Guidance on Evaluating Instrument Based Measures).
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well. (see Guidance on Evaluating Evidence for Appropriate Use Measures).

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

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

1c. For composite performance measures, the following must be explicitly articulated and logical

1c1. 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

1c2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and

1c3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

[3] These criteria apply to all types of measures except for cost, resource use, and efficiency measures. See Evaluation Criteria for Cost and Resource Use Measures for criteria specific to those types of measures.

2. Scientific Acceptability of Measure Properties: Reliability and Validity

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

2a. Reliability

2a1. The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability.

2a2. Reliability testing demonstrates that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs), reliability must be demonstrated for the data element level as well as for the computed performance score. For **composite performance measures**, reliability must be demonstrated for the computed performance score.

2b. Validity

2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures** (including **PRO-PMs)** validity must be demonstrated for the data element level as well as for the computed performance score. For **composite performance measures**, validity must be demonstrated for the computed performance score by the time of endorsement maintenance; if empirical testing of the computed performance score 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 (via either empirical testing or face validity) .

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

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).

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care, and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment. (See section on Risk Adjustment for Social Risk Factors)

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance;

OR

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

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

2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate the following:

2c1. 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
2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

3. Feasibility:
Extent to which the specifications, including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

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

3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3c. Demonstration that the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient-reported data, patient confidentiality, costs associated with fees/licensing for proprietary measures or elements such as risk model, grouper, instrument) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

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 to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Use (***must-pass*** for maintenance of endorsement)
4a1. Accountability and Transparency
Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4a2. Feedback on the measure by those being measured or others is demonstrated when:
1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data
2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation
3) this feedback has been considered when changes are incorporated into the measure

4b. Usability (***NOT must-pass*** for maintenance of endorsement)[4]

4b1. Improvement

[4] For the present, this subcriterion is not considered must-pass due to concerns that data may not indicate improvement, even when improvement has occurred, and because evidence of unintended negative consequences often is unavailable.

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.[5] 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.

4b2. 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).

5. Comparison to Related or Competing Measures
If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5a. The measure specifications are harmonized with related measures
OR
the differences in specifications are justified.

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

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

NQF Measure Evaluation Criteria and Guidance

Guidance on Evaluating Importance to Measure and Report

1. Importance to Measure and Report: Evidence and Performance Gap
Extent to which the specific measure focus is evidence-based and important to making significant gains in healthcare quality where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria*.**

1a. Evidence to Support the Measure Focus
See Algorithm 1 and Table 2 for guidance on how to rate this subcriterion:
High ☐ Moderate ☐ Low ☐ Insufficient ☐
OR
For outcome measures, see Algorithm 1 for guidance on how to rate this subcriterion: Pass ☐ No Pass ☐

[5] 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. 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.

The measure focus is evidence-based, demonstrated as follows:

- Outcome: Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 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 that the measured structure leads to a desired health outcome.
- Efficiency: evidence is required for the quality component but not required for the resource use component. (Measures of efficiency combine the concepts of resource use and quality.)
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful (see Table 13 under Guidance on Evaluating Instrument Based Measures).
- Process Measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well. (see Guidance on Evaluating Evidence for Appropriate Use Measures).

  - See **Table** 1 for examples.

  - Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

  - The preferred system for grading the evidence are Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and or modified GRADE.

  - Evidence for specific timeframes or thresholds included in a measure should be presented. If evidence is limited, then literature regarding standard norms would be considered.

  - Examples of evidence to demonstrate that the target population for patient-reported measures values the measured outcome, process, or structure and finds it meaningful includes, but is not limited to, patient input in the development of the instrument, survey, or tool; focus group input regarding the value of the performance measure derived from the instrument/survey/tool.

  - Current requirements for structure and process measures (i.e., a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured structure/process leads to a desired health outcome) also apply to patient-reported structure/process measures.

  - Domains of patient-reported outcomes include health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.

  - Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For evidence, if the steward/developer attests that the evidence for a measure has not changed since its previous endorsement evaluation, there is a decreased emphasis on evidence, meaning that the standing committee may accept the prior evaluation of this criterion without further discussion or need for a vote. This applies only to measures that previously passed the evidence criterion without an exception. If a measure was granted an evidence exception, the evidence for that measure must be revisited.

1b. Performance Gap Use Table 3 to rate this subcriterion. **High**☐ **Moderate**☐ **Low**☐ **Insufficient** ☐

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

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

  - When assessing measure performance data for Performance Gap (1b), the following factors should be considered:
    - distribution of performance scores
    - number and representativeness of the entities included in the measure performance data
    - data on disparities
    - size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.

  - 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.

  - Performance Gap (i.e., opportunity for improvement) should be considered differently for some outcome measures such as mortality and patient safety events, where it may be appropriate to continue measurement even with low event rates. Process measures can reasonably reach near 100% performance with little opportunity for additional meaningful gains. For mortality and adverse events measures, however, it is less clear how low is attainable.

For all measures that use ICD-10 coding: For Fall 2017 and CY2018 submissions, performance gap can be based on literature and/or data based on ICD-9 or ICD-10 coding. For CY2019 and beyond, gap information must be based on ICD-10 coded data. If lack of availability of ICD-10 coded data prohibits adherence to this requirement, NQF may grant a grace period for provision of ICD-10 based data. This must be determined on a case-by-case basis prior to the intent-to-submit deadline. If the grace period is granted, CY2018 testing requirements apply.

  - For maintenance of endorsement**:** If a measure is found to be "topped out" (i.e., does not meet criteria for opportunity for improvement (1b)), the measure will be considered for inactive endorsement with reserve status only. The measure must meet all other criteria, otherwise the measure should not be endorsed. See Inactive Endorsement with Reserve Status policy.

  - For maintenance of endorsement**:** Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For performance gap, there is increased emphasis on current performance and opportunity for improvement. Measure stewards are expected to provide current performance data. If limited data are available (e.g., use is voluntary), data from the literature can be considered.

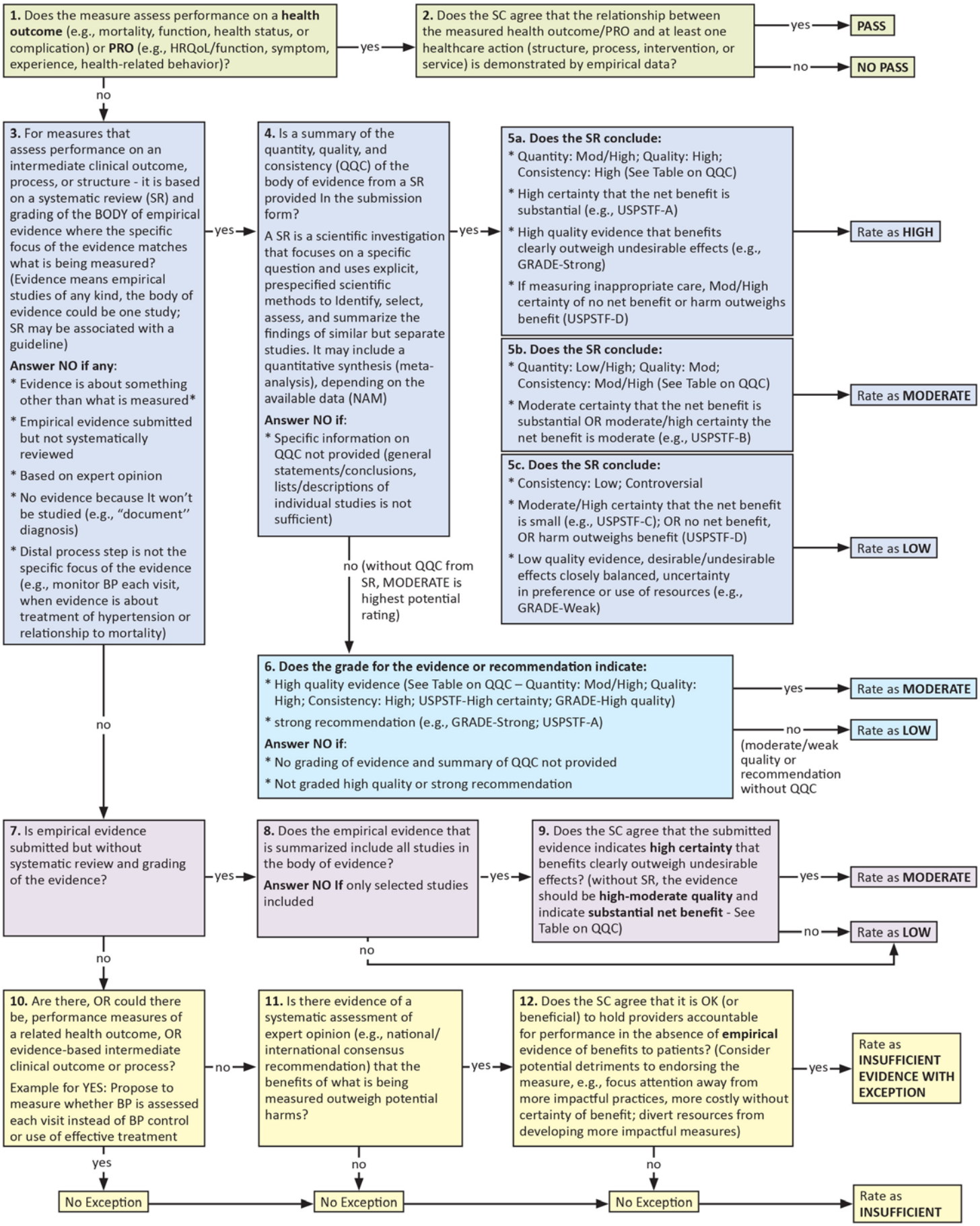
1c. For composite performance measures, the following must be explicitly articulated and logical:
Use Table 3 to rate criterion: **High**☐ **Moderate**☐ **Low**☐ **Insufficient**

1c1. 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
1c2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
1c3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

Table 1. Evidence to Support the Focus of Measurement

Type of Measure	Evidence	Example of Measure Type and Evidence to Be Addressed
Outcome, including Patient-Reported Outcome In some situations, resource use may be considered a proxy for a health state (e.g., hospitalization may represent deterioration in health status).	Empirical data demonstrate a relationship between the outcome to at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.	**#0230** Acute myocardial Infarction 30-day mortality • Survival is a goal of seeking and providing treatment for AMI. • **Data** show that interventions such as aspirin or reperfusion leads to decreased mortality/ increased survival **#0171** Acute care hospitalization (risk-adjusted) [of home care patients] • Improvement or stabilization of condition to remain at home is a goal of seeking and providing home care services. • **Data** show that actions such as medication reconciliation or care coordination) leads to decreased hospitalization of patients receiving home care services **#0166** HCAHPS experience with communication with doctors (assuming demonstration this is of value to patients) • **Data** show that healthcare practices such as response time, respect, attention, or explanation leads to better experience with physician communication
Intermediate Clinical Outcome	Quantity, quality, and consistency of a body of evidence that the measured intermediate clinical outcome leads to a desired health outcome.	**#0059** Hemoglobin A1c management [A1c > 9] • **Evidence** that hemoglobin A1c level leads to health outcomes (e.g., prevention of renal disease, heart disease, amputation, mortality)
Process	Quantity, quality, and consistency of a body of evidence that the measured healthcare process leads to desired health outcomes in the target population with benefits that outweigh harms to patients. Specific drugs and devices should have FDA approval for the target condition. If the measure focus is on inappropriate use, then quantity, quality, and consistency of a body of evidence that the measured healthcare process does *not* lead to desired health outcomes in the target population.	**#0551** ACE inhibitor/Angiotensin receptor blocker (ARB) use and persistence among members with coronary artery disease at high risk for coronary events • **Evidence** that use of ACE-I and ARB results in lower mortality and/or cardiac events **#0058** Inappropriate antibiotic treatment for adults with acute bronchitis • **Evidence** that antibiotics are not effective for acute bronchitis
Structure	Quantity, quality, and consistency of a body of evidence that the measured healthcare structure leads to desired health outcomes with benefits that outweigh harms (including evidence for the link to effective care processes and the link from the care processes to desired health outcomes).	**#0190** Nurse staffing hours • **Evidence** that higher nursing hours result in lower mortality or morbidity, or lead to provision of effective care processes (e.g., lower medication errors) that lead to better outcomes

Algorithm 1. Guidance for Evaluating the Clinical Evidence



NOTE: Submissions for instrument-based measures that are based on patient report must include information to demonstate that the target population values the measured structure, process, or outcome and finds it meaningful. If not demonstrated, then rating should be INSUFFICIENT.

Table 2. Evaluation of Quantity, Quality, and Consistency of Body of Evidence for Structure, Process, and Intermediate Outcome Measures

Definition/ Rating	Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence
Definition	Total number of studies (not articles or papers)	Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors[6] including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events)	Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence
High	5+ studies[7]	Randomized controlled trials (RCTs) providing direct evidence for the specific measure focus, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction and similar in magnitude across the preponderance of studies in the body of evidence
Moderate	2-4 studies	• Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect OR • RCTs without serious flaws that introduce bias, but with either indirect evidence or imprecise estimate of effect	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction across the preponderance of studies in the body of evidence, but may differ in magnitude If only one study, then the estimate of benefits greatly outweighs the estimate of potential harms to patients (one study cannot achieve high consistency rating)

[6] ***Study designs*** that affect certainty of confidence in estimates of effect include randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders. ***Study flaws*** that may bias estimates of effect include lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes. ***Imprecision*** with wide confidence intervals around estimates of effects can occur in studies involving few patients and few events. ***Indirectness*** of evidence includes indirect comparisons (e.g., two drugs compared to placebos rather than head-to head); and differences between the population, intervention, comparator interventions, and outcome of interest and those included in the relevant studies. Source: Guyatt GH, Oxman AD, Kunz R, et al., What is "quality of evidence" and why is it important to clinicians?, *BMJ*, 2008;336(7651):995-998.

[7] The suggested number of studies for rating levels of quantity is considered a general guideline.

Definition/ Rating	Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence
Low	1 study	• RCTs with flaws that introduce bias OR • Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations	• Estimates of clinically/practically meaningful benefits and harms to patients differ in both direction and magnitude across the preponderance of studies in the body of evidence OR • wide confidence intervals prevent estimating net benefit If only one study, then estimate of benefits do not greatly outweigh harms to patients
Insufficient to Evaluate	• No empirical evidence OR • Only selected studies from a larger body of evidence	• No empirical evidence OR • Only selected studies from a larger body of evidence	No assessment of magnitude and direction of benefits and harms to patients

Table 3: Generic Scale for Rating Performance Gap and Quality Construct Subcriteria (1b, 1c, 2c)

Rating	Definition
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

Guidance on Evaluating Scientific Acceptability of Measure Properties

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

2a. Reliability See **Algorithm 2** for guidance on how to rate this subcriterion:
High☐ **Moderate**☐ **Low**☐ **Insufficient**☐

- **NOTE on Algorithm 2**: This algorithm provides general guidance on how to rate measures for reliability; however, it may not be completely applicable for all measures. For example, instrument-based measures require reliability testing at both the data element and measure score levels, but this special case this isn't called out in the algorithm.

2a1. The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability.

- 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.

- All measures that use the ICD classification system must use ICD-10-CM.

- eCQMs should be specified using the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC).

- Specifications for instrument-based measures also include the specific instrument (e.g., PROM(s)); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the performance measure results.

- Specifications for composite performance measures include component measure specifications (unless individually endorsed); aggregation and weighting rules; handling of missing data; standardizing scales across component measures; required sample sizes.

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. However, there is no change in the evaluation of the current specifications.

2a2. Reliability testing demonstrates that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures (**including PRO-PMs), reliability must be demonstrated for the data element level as well as for the computed performance score. For **composite performance measures**, reliability must be demonstrated for the computed performance score.

- 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). See Table 4 for guidance regarding testing requirements by measure type and the preferred scope of testing expected at the time of endorsement maintenance.

- Testing must be conducted for the measure as specified (e.g., all relevant levels of analysis, using applicable data sources, care settings, patients, providers, etc.). If more than one measure is included under one NQF number, each measure must be tested per NQF evaluation requirements. **If more than one level of analysis is specified, testing must be conducted for each level separately.**

- Testing at the level of data elements requires that all critical data elements be tested (not just agreement of one final overall computation for all patients). At a minimum, the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately.

- For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred.

- For eCQMs:
  - Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid (and this must be demonstrated empirically).
  - If sufficient data are available for testing, testing of reliability and validity at the score level is encouraged (in addition to testing at the data element level).
  - If a developer is testing an eCQM using any type of normalized EHR clinical data (e.g. from multiple EHR sources), NQF requires, at a minimum, supporting information of what schemas are included in the normalized

data set and how they are calculated by the measure logic (i.e., what fields have been normalized and how, including any considerations of how this may affect the measure).

- **Beginning Summer 2019**:
  - Reliability of unstructured data must be demonstrated at the data element level.
  - If data element testing is not possible, justification is required and must be accepted by the Standing Committee.
  - If sufficient data are available for testing, testing of reliability and validity at the score level is encouraged in addition to required data element testing.

- Samples used for testing:
  - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
  - The sample should represent the variety of entities whose performance will be measured. The 2010 Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
  - The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
  - When possible, units of measurement and patients within units should be randomly selected.

- For some measure types, separate reliability testing of the data elements is not required if empirical validity testing of the data elements is conducted (and results are adequate).

- Prior evidence of reliability of data elements for the data type specified in the measure (e.g., hospital claims) can be used as evidence for those data elements. Prior evidence could include published or unpublished testing that: includes the same data elements, uses the same data type (e.g., claims, chart abstraction, etc.), and is conducted on a sample as described above (i.e., representative, adequate numbers, and randomly selected, if possible).

- For measures that use ICD-10 coding: For Fall 2017 and CY2018 submissions, submit updated ICD-10 reliability testing if available; if not, testing based on ICD-9 coding will suffice. For CY2019 and beyond, reliability testing must be based on ICD-10 coded data. If lack of availability of ICD-10 coded data prohibits adherence to this requirement, NQF may grant a grace period for provision of ICD-10 based testing. This must be determined on a case-by-case basis prior to the intent-to-submit deadline. If the grace period is granted, CY2018 testing requirements apply.

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For reliability testing, if no new testing information is presented, the Committee may accept the prior evaluation of the testing results without further discussion or need for a vote, **as long as the previous testing conforms to current requirements**.

- NQF's Scientific Methods Panel will provide NQF standing committees with evaluations and ratings of reliability and validity for new complex measures and for previously endorsed complex measures with updated testing (i.e., those with new information for testing, including additional statistics or testing based on a different timeframe, data source, etc.). For the purposes of Scientific Methods Panel evaluation, complex measures are defined as outcome measures, including intermediate clinical outcomes; instrument-based measures (e.g., patient-report outcome-based performance measures); cost/resource use measures; efficiency measures (those combining concepts of resource use and quality); and composite measures.

2b. Validity See **Algorithm 3** for guidance on how to rate this subcriterion:
High ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

- **NOTE on Algorithm 3**: This algorithm provides general guidance on how to rate measures for validity; however, it may not be completely applicable for all measures. For example, instrument-based measures require validity

testing at both the data element and measure score levels, but this special case this isn't called out in the algorithm.

2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures** (including **PRO-PMs),** validity must be demonstrated for the data element level as well as for the computed performance score. For **composite performance measures**, validity must be demonstrated for the computed performance score by the time of endorsement maintenance; if empirical testing of the computed performance score 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 (via either empirical testing of the data elements or measure score or via face validity).

- Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures).

- Testing must be conducted for the measure as specified (e.g., all relevant levels of analysis, using applicable data sources, care settings, patients, providers, etc.). If more than one measure is included under one NQF number, each measure must be tested per NQF evaluation requirements**. If more than one level of analysis is specified, testing must be conducted for each level separately.**

- Testing at the level of data elements requires that all critical data elements be tested (not just agreement of one final overall computation for all patients). At a minimum, the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately.

- If presenting score-level validation (typically via construct validity or known-groups analysis) the following should be included:
    - Narrative describing the hypothesized relationships
    - Narrative describing why examining these relationships (e.g., correlating measures) would validate the measure
    - Expected direction of the association
    - Expected strength of the association
    - Specific statistical tests used (more detail is better)
    - Results of the analysis
    - Interpretation of those results (including how they related to the hypothesis and whether they have helped to validate the measure)

- For eCQMs:
    - Beginning September 30, 2017, all respecified measure submissions for use in federal programs (previously known as "legacy" eCQMs) will be required to the same evaluation criteria as respecified measures – the "BONNIE testing only" option will no longer meet endorsement criteria for validity. NOTE that testing (e.g., using BONNIE) is still needed to confirm that the measure logic works as expected and that value sets are included in the VSAC.
    - Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid (and this must be demonstrated empirically).
    - If sufficient data are available for testing, testing of validity at the score level is encouraged (in addition to testing at the data element level).
    - If a developer is testing an eCQM using any type of normalized EHR clinical data (e.g. from multiple EHR sources), NQF requires, at a minimum, supporting information of what schemas are included in the normalized

data set and how they are calculated by the measure logic (i.e., what fields have been normalized and how, including any considerations of how this may affect the measure).
  - **Beginning Summer 2019**:
    - Validity must be demonstrated at the data element level.
    - If data element testing is not possible, justification is required and must be accepted by the Standing Committee.
    - Face validity by itself is not sufficient for eCQMs, whether new or maintenance.
    - If sufficient data are available for testing, testing of reliability and validity at the score level is encouraged in addition to required data element testing.

- Samples used for testing:
  - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
  - The sample should represent the variety of entities whose performance will be measured. The 2010 Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
  - The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
  - When possible, units of measurement and patients within units should be randomly selected.

- Prior evidence of validity of data elements for the data type specified in the measure (e.g., hospital claims) can be used to demonstrate validity for those data elements. Prior evidence could include published or unpublished testing that: includes the same data elements, uses the same data type (e.g., claims, chart abstraction, etc.), and is conducted on a sample as described above (i.e., representative, adequate numbers, and randomly selected, if possible).

- Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed. See Table 4 for guidance regarding testing requirements by measure type and the preferred scope of testing expected at the time of endorsement maintenance.

- For all measures that use ICD-10 coding (see Guidance for Measures Using ICD-10 Coding): Beginning with Fall 2017 submissions, updated validity testing must be submitted:
  - Submit updated empirical validity testing on the ICD-10 specified measure, if available
  - **OR** face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of quality
  - **OR** face validity of the ICD-10 coding scheme plus score-level empirical validity testing based on ICD-9 coding
  - **OR** face validity of the ICD-10 coding scheme plus data element level validity testing based on ICD-9 coding, with face validity of the measure score as an indicator of quality due at annual update
  - **For CY2019 and beyond**, validity testing must be based on ICD-10 coded data; if providing face validity, both face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of quality is required. If lack of availability of ICD-10 coded data prohibits adherence to this requirement for maintenance measures, NQF may grant a grace period for provision of ICD-10 based testing. This must be determined on a case-by-case basis prior to the intent-to-submit deadline. If the grace period is granted, CY2018 testing requirements apply.

- For maintenance of endorsement: For non-eCQMS, empirical validity testing is expected at time of maintenance review; if not possible, justification is required. For eCQMs, empirical testing of the data elements will be required **as of Summer 2019**. If data element testing is not possible, justification is required and must be accepted by the Standing Committee.

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For validity, there is less emphasis on the criterion if additional testing information has not been presented, and the Committee may accept the prior evaluation of this subcriterion without further discussion and vote, **as long as the previous testing conforms to current requirements**. For outcome measures, the committee discusses questions related to adjustment for social risk factors, even if no change in testing is presented.

- NQF's Scientific Methods Panel will provide NQF standing committees with evaluations and ratings of reliability and validity for new complex measures and for previously endorsed complex measures with updated testing (i.e., those with new information for testing, including additional statistics or testing based on a different timeframe, data source, etc.). For the purposes of Scientific Methods Panel evaluation, complex measures are defined as outcome measures, including intermediate clinical outcomes; instrument-based measures (e.g., patient-report outcome-based performance measures); cost/resource use measures; efficiency measures (those combining concepts of resource use and quality); and composite measures.

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

- 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.

- Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions

2b3. For outcome measures and other measures when indicated (e.g., resource use):
- an evidence-based risk-adjustment strategy is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care, and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment. (See section on Risk Adjustment for Social Risk Factors)

- Risk factors that influence outcomes should not be specified as exclusions.

- In July 2017, the NQF Board of Directors reviewed findings from the 2-year SDS Trial and agreed to continue suspension of the policy that prohibits use of social risk factors in risk-adjustment approaches. Therefore, for the present, risk-adjusted measures submitted to NQF for evaluation **may include** both clinical and social risk factors in the risk adjustment models. See section on risk adjustment for social risk factors.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance;
OR
there is evidence of overall less-than-optimal performance.

- 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 vs. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 vs. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

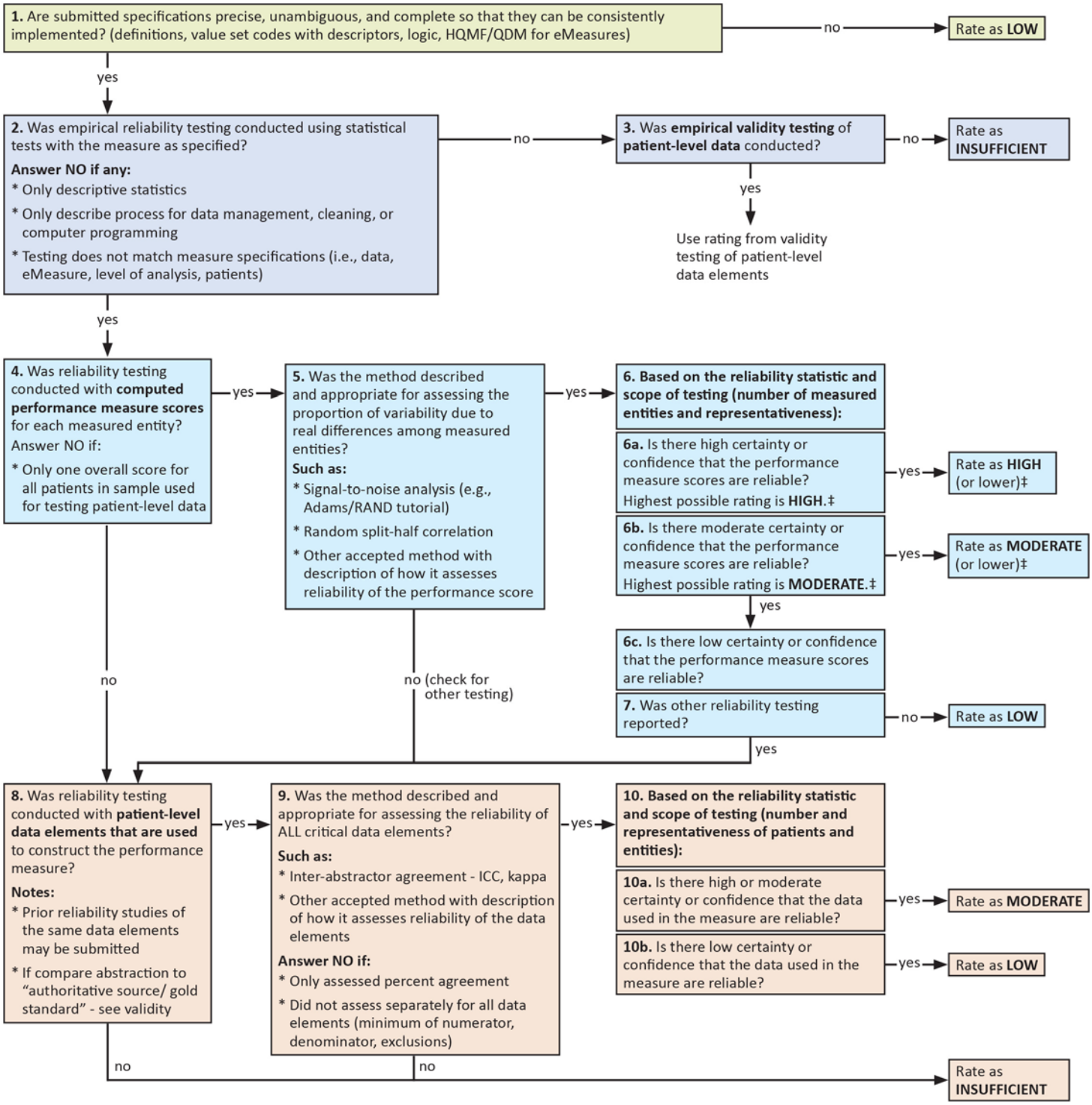
2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate the following: Use Table 3 to rate this subcriterion. **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient** ☐

2c1. 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

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

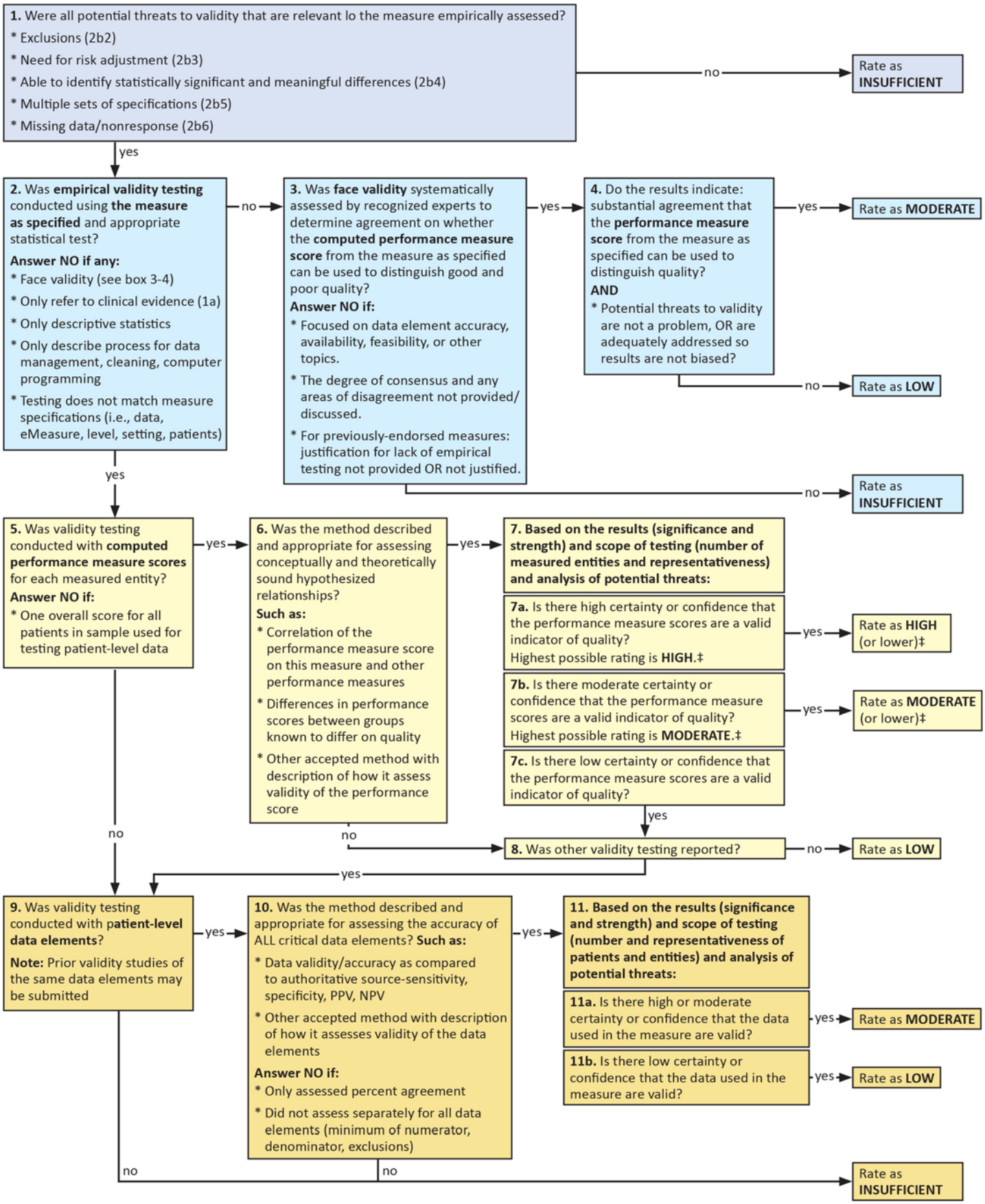
(*if not conducted or results not adequate, justification must be submitted and accepted)*

Algorithm 2. Guidance for Evaluating Reliability



‡ This is the highest possible rating, but it may be lower, depending on the strength of data element testing results. If data element testing is provided, these results must also be considered.

Algorithm 3. Guidance for Evaluating Validity



‡ This is the highest possible rating, but it may be lower, depending on the strength of data element testing results. If data element testing is provided, these results must also be considered.

Table 4. Testing Requirements by Measure Type and Preferred Scope of Testing Expected at the Time of Evaluation for Endorsement Maintenance

Measure type	Requirements for reliability testing	Requirements for validity testing
Instrument-based measures	BOTH data element and score-level testing	BOTH data element and score-level testing
Composite measures	Score-level testing of the composite measure score; testing of the components is not sufficient.	Score-level testing of the composite measure score is desired. At initial endorsement only, empirical or face validity testing of the components OR face validity of the composite is acceptable.
eCQMs	All eCQMs must be tested using the eCQM specifications. These must use the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC).. Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid (and this must be tested empirically). Thus, testing for elements that are not included in structured data fields should be tested at the data element level.	All eCQMs must be tested using the eCQM specifications. These must use the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC).Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid (and this must be tested empirically). Thus, testing for elements that are not included in structured data fields should be tested at the data element level. Empirical testing is expected, **and as of Summer 2019**, data element validation will be required unless justification is provided/accepted. Face validity alone will not be sufficient. Use of a simulated data set (e.g. BONNIE) is no longer accepted for testing validity of data elements.
Cost and Resource Use Cost and Resource Use Measure Evaluation Criteria	EITHER data element or score-level testing	Validity is considered in the context of measure intent and threats to validity based on these cost measure-specific components: • Attribution approach • Cost categories • Approach to outliers • Impact of Carve Outs EITHER data element or score-level testing; face validity not accepted for maintenance measures unless justification provided/accepted
All others	EITHER data element or score-level testing	EITHER data element or score-level testing; face validity not accepted for maintenance measures unless justification provided/accepted; if data element validity is demonstrated, additional reliability testing is not required

Ongoing testing and evaluation of the measure should be performed to understand how a measure is being used in the field. The remainder of this table provides guidance regarding the preferred scope of testing for measures undergoing maintenance evaluation. Note that this guidance does not supersede testing requirements for specific measures types as shown above.

	First Maintenance Evaluation	Subsequent Maintenance Evaluations
Reliability	Measure In Use • Analysis of data from entities whose performance is measured • Reliability of measure scores Measure Not in Use • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that reliability achieved at least a moderate rating
Validity	Measure in Use • Analysis of data from entities whose performance is measured • Validity of measure score for making accurate conclusions about quality • Updated/expanded analysis of threats to validity Measure Not in Use • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that validity achieved at least a moderate rating

Guidance on Evaluating Feasibility

3. Feasibility:
Extent to which the specifications, including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement. Use **Table 5** to rate this subcriterion.

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

3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3c. Demonstration that the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient-reported data, patient confidentiality, costs associated with fees/licensing for proprietary measures or elements such as risk model, grouper, instrument) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

- 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.

- For all eCQMs, a feasibility assessment is required. This feasibility assessment must address the data elements and measure logic and demonstrate that the eCQM can be implemented or that feasibility concerns can be adequately addressed. The feasibility assessment uses a standard score card. BONNIE testing (or some other type of testing)

should be used to demonstrate that the measure logic will work (including demonstration of 100% coverage of the measure logic using simulated data). See section on eCQMs.

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. However, the emphasis on this criterion is the same for both new and previously endorsed measures, as feasibility issues might have arisen for endorsed measures that have been implemented.

Table 5. Generic Scale for Rating Feasibility Criterion

Rating	Definition
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

Guidance on Evaluating Usability and Use

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 to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Use (***must-pass*** for maintenance of endorsement): Use **Table 7** to rate this criterion as **Pass** ☐ **No Pass** ☐

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For Use, there is increased emphasis on the use of the measure, especially use for accountability purposes. There also is an increased emphasis on improvement in results over time and on unexpected findings, both positive and negative.

4a1. Accountability and Transparency

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

- ***Transparency*** is the extent to which performance results about identifiable, accountable entities are *disclosed and available* outside of the organizations or practices whose performance is measured. Maximal transparency is achieved with ***public reporting***, defined as making comparative performance results about identifiable, accountable entities freely available (or at nominal cost) to the public at large (generally on a public website). At a minimum, the data on performance results about identifiable, accountable entities are available to the public (e.g., unformatted database). The capability to verify the performance results adds substantially to transparency.

- Accountability applications are uses 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.

- A 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.

- Measures that are included in finalized rule for federal public reporting programs will be considered publicly reported, even if not yet implemented.

4a2. Feedback on the measure by those being measured or others is demonstrated when:

1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data
2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation
3) this feedback has been considered when changes are incorporated into the measure

- For measures not previously endorsed, a plan for how feedback will be collected and used should be discussed.

- Information to address this subcriterion include:

    - For (1): describing how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation; describing how many and which types of measured entities and/or others were given this information (if only a sample of measured entities were included, describing the full population and how the sample was selected); describing the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

    - For (2): summarizing the feedback obtained on measure performance and implementation from the measured entities and others and how that feedback was obtained. This could also include the amount of feedback, which stakeholders had substantial feedback, etc.

    - For (3): describing how the feedback has been considered when developing or revising the measure specifications or implementing the measure, including whether the measure was modified and why or why not it was modified based on feedback received.

4b. Usability (***NOT must-pass*** for maintenance of endorsement)[8]
Use **Table 6** and **Table 7** to rate this criterion as **High**☐ **Moderate**☐ **Low**☐ **Insufficient**☐

- Under NQF's revised approach to the evaluation of currently endorsed measures, there is a shift in emphasis for several of the evaluation criteria/subcriteria. For Usability, there is increased emphasis on improvement in results over time and on unexpected findings, both positive and negative.

4b1. Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

- 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. 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.

4b2. 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).

[8] For the present, this subcriterion is not considered must-pass due to concerns that data may not indicate improvement, even when improvement has occurred, and because evidence of unintended negative consequences often is unavailable.

- Information for subcriteria 4b1 and 4b2 may be obtained via literature, feedback to NQF, and from measure stewards/developers during the submission process.

Table 6: Generic Scale for Rating Usability and Use Criterion

Rating	Definition
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

Table 7. Key Questions for Evaluating Usability and Use

Subcriteria	Key Questions	Suitable for Endorsement?
4a, 4b	• Are the subcriteria met? (4a1—accountability/transparency, 4a2—feedback on measures, 4b1—improvement, and 4b2—benefits outweigh any unintended consequences)	If Yes, then the Usability and Use criterion is met, and if the other criteria (Importance to Measure and Report, Scientific Acceptability of Measure Properties, Feasibility) are met, then the measure is **suitable for endorsement**

Subcriteria	Key Questions	Suitable for Endorsement?
4a1. Accountability/Transparency	• Is it an initial submission with a credible plan for implementation in an accountability application? • Is the measure used in at least one accountability application within three years? • Are the performance results publicly reported within six years (or the data on performance results are available)? If any of the above answers are "No": • What are the reasons (e.g., developer/steward, external factors)? • Is there a credible plan for implementation and public reporting?	If 4a1 is not met when initially submitted, then the Usability and Use criterion is not met, but the measure **may or not be suitable for endorsement** depending on an assessment of the following: • timeframe (initial submission, three years, six years, or longer); • reasons for lack of use in accountability application/public reporting (4a1); • credibility of plan for implementation for accountability/public reporting (4a1); • strength of the measure in terms of the other three criteria (Importance to Measure and Report, Scientific Acceptability of Measure Properties, and Feasibility); and • strength of competing and related measures to drive improvement. Exceptions to the timeframes for accountability and public reporting (4a1) judgment and supporting rationale.
4a2. Feedback on the measure by those being measured or others	• Does the information demonstrate feedback by those being measured or others?	Not a "must-pass" criterion for *initial* endorsement but is a must-pass criterion for maintenance of endorsement.

Subcriteria	Key Questions	Suitable for Endorsement?
4b1. Improvement	• Is it an initial submission with a credible rationale for improvement? • Has improvement been demonstrated (performance trends, numbers of people receiving high-quality, efficient healthcare)? If any of the above answers are "No": • What are the reasons? • Is there a credible rationale describing how the performance results could be used to further the goal of facilitating high-quality, efficient healthcare for individuals or populations? • Is the measure used in quality improvement programs?	If 4b1 is not met, then the Usability and Use criterion is not met, but the measure **may or not be suitable for endorsement** depending on an assessment of the following: • timeframe (initial submission, three years, six years, or longer); • reasons for lack of improvement (4b1); • credibility of rationale for improvement (4b1); • strength of the measure in terms of the other three criteria (Importance to Measure and Report, Scientific Acceptability of Measure Properties, and Feasibility); and • strength of competing and related measures to drive improvement. Exceptions to the demonstration of improvement (4b1) require judgment and supporting rationale.
4b2. Unintended negative consequences	• Is there **evidence** that unintended negative consequences to individuals or populations outweigh the benefits? For most measures, this will not be applicable and will not be a factor in whether a measure is recommended.	If Yes, then the Usability and Use criterion is not met and the measure is **not suitable for endorsement** regardless of evaluation of 4a1, 4a2, and 4b1.

Guidance on Evaluating Related and Competing Measures

5. Comparison to Related or Competing Measures
If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.
See Table 8, Table 9, Table 10, and Figure 1.

5a. The measure specifications are harmonized with related measures
OR
the differences in specifications are justified.

- Measure **harmonization** refers to the standardization of specifications for related measures with the same measure focus (e.g., *influenza immunization* of patients in hospitals or nursing homes); related measures with the same target population (e.g., eye exam and HbA1c for *patients with diabetes*); or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are

justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

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

Table 8. Related versus Competing Measures

	Same concepts for measure focus—target process, condition, event, outcome	**Different concepts for measure focus—target process, condition, event, outcome**
Same target patient population	**Competing measures—Select best measure** from competing measures or justify endorsement of additional measure(s).	**Related measures—Harmonize** on target patient population or justify differences.
Different target patient population	**Related measures—Combine** into one measure with expanded target patient population or justify why different harmonized measures are needed.	Neither harmonization nor competing measure issue

Figure 1. Addressing Competing Measures and Harmonization of Related Measures in the NQF Evaluation Process

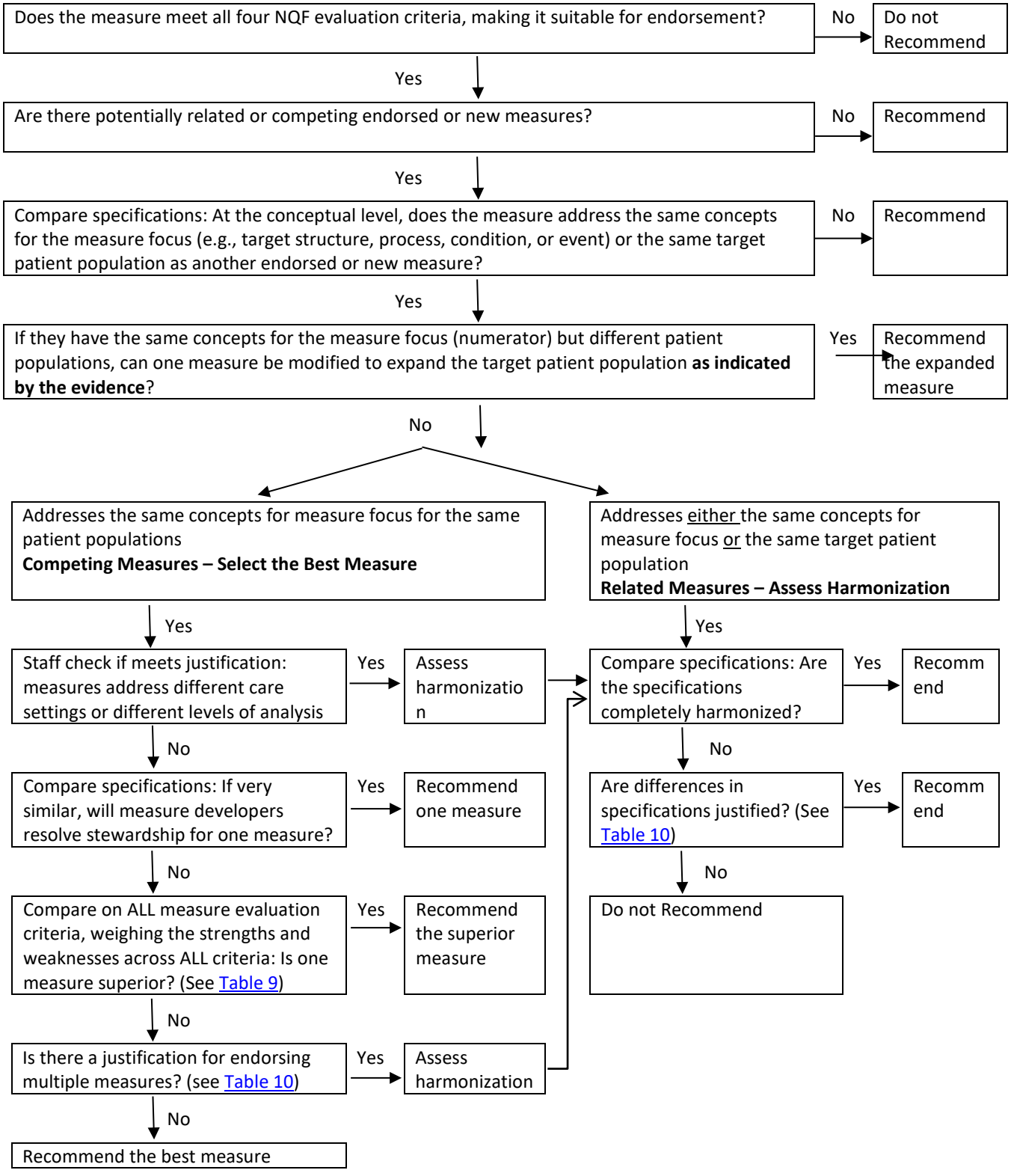


Table 9. Evaluating Competing Measures for Superiority or Justification for Multiple Measures

Steps	Evaluate Competing Measures
1. Determine if need to compare measures for superiority	Work through the steps in the algorithm (Figure 1) to determine if need to evaluate competing measures for superiority (i.e., two or more measures address the same concepts for measure focus for the same patient populations)
2.Assess Competing Measures for Superiority by weighing the strengths and weaknesses across ALL NQF evaluation criteria	**Because the competing measures have already been determined to have met NQF's criteria for endorsement, the assessment of competing measures must include weighing the strengths and weaknesses across ALL the criteria and involves more than just comparing ratings.** (For example, a decision is not based on just the differences in scientific acceptability of measure properties without weighing the evaluation of importance to measure and report, usability, and feasibility as well.) **Evidence, Performance Gap—Importance to Measure and Report:** Competing measures generally will be the same in terms of the evidence for the focus of measurement (1a). However, due to differences in measure construction, they could differ on performance gap (1b). • Compare measures on opportunity for improvement (1b) **Reliability and Validity—Scientific Acceptability of Measure Properties:** • Compare evidence of reliability (2a1-2a2) • Compare evidence of validity, including threats to validity (2b1-2b6) Untested measures cannot be considered superior to tested measures because there would be no empirical evidence on which to compare reliability and validity. However, a new measure, when tested, could ultimately demonstrate superiority over an endorsed measure and the NQF endorsement maintenance cycles allow for regular submission of new measures. Compare and identify differences in specifications **All else being equal on the criteria and subcriteria**, the preference is for: • Measures specified for the broadest application (target patient population as indicated by the evidence, settings, level of analysis) • Measures that address disparities in care when appropriate **Feasibility:** • Compare the ease of data collection/availability of required data **All else being equal on the criteria and subcriteria**, the preference is for: • Measures based on data from electronic sources • Clinical data from EHRs • Measures that are freely available **Usability and Use:** • Compare evidence of the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement. **All else being equal on the criteria and subcriteria**, the preference is for: • Measures used in at least one accountability application • Measures with the widest use (e.g., settings, numbers of entities reporting performance results) • Measures for which there is evidence of progress towards achieving high-quality efficient healthcare for individuals or populations • The benefits of the measure outweigh any unintended negative consequences to individuals or populations **After weighing the strengths and weaknesses across ALL criteria, identify if one measure is clearly superior and provide the rationale based on the NQF criteria.**

Steps	Evaluate Competing Measures
3.If a competing measure does not have clear superiority, assess justification for multiple measures	**If a competing measure does not have clear superiority, is there a justification for endorsing multiple measures? Does the added value offset any burden or negative impact?** **Identify the value of endorsing competing measures** Is an additional measure necessary? • to change to EHR-based measurement; • to have broader applicability (if one measure cannot accommodate all patient populations; settings, e.g., hospital, home health; or levels of analysis, e.g., clinician, facility; etc.); • to increase availability of performance results (if one measure cannot be widely implemented, e.g., if measures based on different data types increase the number of entities for whom performance results are available) **Note:** Until clinical data from electronic health records (EHRs) are widely available for performance measurement, endorsement of competing measures based on different data types (e.g., claims and EHRs) may be needed to achieve the dual goals of 1) advocating widespread access to performance data and 2) migrating to performance measures based on EHRs. EHRs are the preferred source for clinical record data, but measures based on paper charts or data submitted to registries may be needed in the transition to EHR-based measures. Is an additional measure unnecessary? • primarily for unique developer preferences **Identify the burden of endorsing competing measures** Do the different measures affect interpretability across measures? Does having more than one endorsed measure increase the burden of data collection? **Determine if the added value of endorsing competing measures offsets any burden or negative impact?** • If yes, recommend competing measures for endorsement (if harmonized) and provide the rationale for recommending endorsement of multiple competing measures. Also, identify analyses needed to conduct a rigorous evaluation of the use and usefulness of the measures at the time of endorsement maintenance. • If no, recommend the best measure for endorsement and provide rationale.

Table 10. Sample Considerations to Justify Lack of Measure Harmonization

Related Measures	Lack of Harmonization	Assess Justification for Conceptual Differences	Assess Justification for Technical Differences
Same measure focus (numerator); different target population (denominator)	Inconsistent measure focus (numerator)	The evidence for the measure focus is different for the different target populations so that one measure cannot accommodate both target populations. Evidence should always guide measure specifications.	• Differences in the available data drive differences in the technical specifications for the measure focus. • Effort has been made to reconcile the differences across measures, but important differences remain.
Same target population (denominator); different measure focus (numerator)	Inconsistent target population (denominator) and/or exclusions	The evidence for the different measure focus necessitates a change in the target population and/or exclusions. Evidence should always guide measure specifications.	• Differences in the available data drive differences in technical specifications for the target population. • Effort has been made to reconcile the differences across measures, but important differences remain.
For any related measures	Inconsistent scoring/ computation	The difference does not affect interpretability or burden of data collection. If it does, it adds value that outweighs any concern regarding interpretability or burden of data collection.	The difference does not affect interpretability or burden of data collection. If it does, it adds value that outweighs any concern regarding interpretability or burden of data collection.

Evaluation Criteria for Cost and Resource Use Measures

These criteria were last updated in May 2019.

1. Importance to Measure and Report (Must-Pass)

1a. High Impact

The measure focus addresses a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

AND

1. Opportunity for Improvement (Performance Gap)

Demonstration of resource use or cost problems and opportunity for improvement (i.e., data demonstrating considerable variation cost or resource across providers)

2. Scientific acceptability of the measure properties (Must-pass)

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the cost or resources used to deliver care. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.

2a. Reliability

2a1. The measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability.

- All measures that use the ICD classification system must use ICD-10-CM.
- eCQMs should be specified using the latest industry accepted eCQM technical specifications: health quality measure format (HQMF), Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC)

2a2. Reliability testing demonstrates that the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, and/or that the measure score is precise.

2b. Validity

2b1. The measure specifications are consistent with the measure intent and captures the most inclusive target population.

2b2. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.

- Face validity of the measure score as a performance indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and if it explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor performance. The degree of consensus and any areas of disagreement must be provided/discussed.
- Beginning in CY2019, for measures that use ICD-10 coding, validity testing should be based on ICD-10 coded data; if providing face validity, both face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of performance are required.
- For eCQMs: Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid. As of August 2019, validity testing at the data element level will be required for all eCQMs. However, as with other measures, testing at the level of the performance measure score also is encouraged if data can be obtained from enough measured entities. If data element testing is not possible, justification is required and must be accepted by the Standing Committee.
- For maintenance of endorsement: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b3. Exclusions are supported by the clinical evidence

- Exclusions are supported by the clinical evidence
- There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of the nonclinical exclusions;
- The effect of exclusions on the measure score is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);
- If patient preference (e.g., informed decision making) 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).

–

2b4. An evidence-based risk-adjustment strategy is specified and is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at the start of care, and has demonstrated adequate discrimination and calibration;

OR

A rationale/data support no risk adjustment/stratification.

- Risk factors that influence outcomes should not be specified as exclusions.
- In July 2017, the NQF Board of Directors reviewed findings from the 2-year SDS Trial and agreed to continue suspension of the policy that prohibits use of social risk factors in risk-adjustment approaches. Therefore, for the present, risk-adjusted measures submitted to NQF for evaluation may include both clinical and social risk factors in the risk adjustment models.

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

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

2b7**.** Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2c**.** 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

A rationale/data to justify why stratification is not necessary or not feasible.

3. Feasibility

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

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

3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3c. Demonstration that the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) or elements such as risk model, grouper, instrument) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

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 to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Use (Must Pass)

4a1. Accountability and Transparency

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

4a2. Feedback on the measure by those being measured or others is demonstrated when:

1. those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data
2. those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation
3. this feedback has been considered when changes are incorporated into the measure

4b. Usability

4b1. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high quality, efficient healthcare for individuals or populations.

4b2. 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).

4b3. Data and result detail are maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement, can be deconstructed to facilitate transparency and understanding.

Guidance for Measures Using ICD-10 coding

General Guidance for measures submitted after October 1, 2015:

1. All measure submissions must be specified in ICD-10-CM/PCS. Per the current NQF guidance, measure submissions should also include ICD-9-CM codes with a description of the transition process used including, a crosswalk of ICD-9 to ICD-10 codes, and intent of the submission.
    a. NOTE: Measures that are specified to capture data retrospectively may continue to be specified in ICD-9-CM depending on the look-back period. Some measures may be specified to capture data retrospectively and prospectively and therefore may be specified using both ICD-9 and ICD-10.

2. ICD-9 CM codes should be included in the submission until testing for the ICD-10 specified measure can be provided.

For Fall 2017 and CY2018 submissions:

- Gap can be based on literature and/or data based on ICD-9 or ICD-10 coding
- Submit updated ICD-10 reliability testing if available; if not, testing based on ICD-9 coding will suffice
- Submit updated validity testing
  - Submit updated empirical validity testing on the ICD-10 specified measure, if available
  - **OR** face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of quality
  - **OR** face validity of the ICD-10 coding scheme plus score-level empirical validity testing based on ICD-9 coding
  - **OR** face validity of the ICD-10 coding scheme plus data element level validity testing based on ICD-9 coding, with face validity of the measure score as an indicator of quality due at annual update

For 2019 and beyond: All measure information **must be** based on the ICD-10 specified measure. If lack of availability of ICD-10 coded data prohibits adherence to this requirement for maintenance measures, NQF may grant a grace period for provision of ICD-10 based testing. This must be determined on a case-by-case basis prior to the intent-to-submit deadline. If the grace period is granted, CY2018 testing requirements apply.

Best practices for ICD-10 coding (see full recommendations report)

- Use team of clinical and coding experts to "identify specific areas where questions of clinical comparability exist, evaluate consistency of clinical concepts, and ensure appropriate conversion"
- Determine intent
- Use appropriate conversion tool (not required, but also not sufficient by itself; if using conversion tool, consider both forward and backward mapping)
- Assess for material change (For existing measures undergoing coding updates and maintenance, the extent to which the population identified with the new code set overlaps with that identified in the old code set should be assessed, if possible. Measure sponsors also should assess, if possible, whether the conversion results in rates that are similar within defined tolerances.). Options include:
  - Test using dual-coded data if possible OR
  - Face validity (using the above code-conversion process, including use of clinical/coding experts) OR
  - Criterion validity (if dual-coded data not available) OR
  - Consistency across time (pre/post conversion)
- Solicit stakeholder comments

Guidance on Evaluating eCQMs

Definition of eCQMs (also known as electronic clinical quality measures or eMeasures): A measure that is specified using the industry accepted eCQM technical specifications: health quality measure format (HQMF), the Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets vetted through the National Library

of Medicine's Value Set Authority Center (VSAC). Alternate forms of electronic measure specifications that do not use the accepted industry specifications are not considered eCQMs.[9]

eCQMs must meet all evaluation criteria that are current at the time of initial submission or endorsement maintenance (regardless of meeting prior criteria or prior endorsement status). Algorithm 1 applies to eCQMs. Algorithms 2-3 are somewhat applicable to eCQMs, except that demonstration of data element reliability will be required for unstructured data fields **as of Summer 2019** and data element validation will be required for all eCQMs **as of Summer 2019**. If data element testing is not possible, justification is required and must be accepted by the Standing Committee.

A new eCQM version of an endorsed measure is not considered an endorsed measure until it has been specifically evaluated and endorsed by NQF. An eCQM should be submitted as a separate measure even if the same or similar measure exists. NQF has included eCQMs in the NQF measure numbering system and has linked eCQMs to measures that are based on the same concept.

Requirements for Endorsing eCQMs

The following guidance addresses and updates the criteria for endorsement of eCQMs.

Specifications

- Measure specifications should use latest accepted versions of the following industry eCQM technical specifications: Health Quality Measure Format (HQMF), Quality Data Model (QDM), and Clinical Quality Language (CQL). Output from the CMS Measure Authoring Tool (MAT) ensures that the measure uses these technical specifications; however, the MAT is not required to produce HQMF.

- Value sets.
    - All eCQMs submitted to NQF must have published value sets within the VSAC as part of the measure.
    - If an eCQM does not have a published value set, then the measure developer must look to see if there is a published value set that aligns with the proposed value set within its measure.
    - If such a published value set does not exist, then the measure developer must demonstrate that the value set is in draft form and is awaiting publication to VSAC.

Each submitted eCQM undergoes a technical review by NQF staff before going to the Standing Committee for evaluation. For this technical review, NQF staff assess that the measure uses the industry accepted eCQM technical specifications; determine if value sets have been vetted through the VSAC; reviews the feasibility of each data element; and make sure the measure logic has been adequately unit tested using a simulated data set.

Feasibility Assessment

- A feasibility assessment (i.e., scorecard), as originally described in the eMeasure Feasibility Assessment report, is required for all eCQMs. The feasibility assessment includes a scorecard to addresses the data elements and an assessment of the measure logic against a simulated data set. All eCQMs should use the latest NQF Feasibility Scorecard that is available. For assessing measure logic, HTML output from the

[9]NQF accepts measures that use EHRs as a data source and that are tested in EHRs (abstraction or local programming) but are not specified and tested with HQMF specifications. These measures, without HQMF specifications, are not considered eCQMs and will be evaluated as traditional measures against the NQF criteria.

CMS Bonnie tool can be used. Alternative unit testing results are acceptable, provided they also demonstrate 100% coverage of the measure logic using simulated data.

Testing for Reliability and Validity

To be considered for NQF endorsement, all eCQMs must be tested empirically using the HQMF specifications. **Beginning Summer 2019**, data element validation will be required for all eCQMs (demonstration of score-level validation is also encouraged). For eCQMs based solely on structured data fields, reliability testing will not be required **if** data element validation is demonstrated. If data element testing is not possible, justification is required and must be accepted by the Standing Committee.

- The minimum requirement is testing in **EHR systems from more than one EHR vendor**. Developers should test on the number of EHR systems they feel appropriate. It is highly desirable that measures are tested in systems from multiple vendors.
- In the description of the sample used for testing, indicate how the eCQM specifications were used to obtain the data.
- eCQMs specified in older HQMF releases that have previously been endorsed do not need to be retested for maintenance. They may, however, need to be updated to accommodate variations in the most current HQMF release. All newly developed measures should be tested using the most current eCQM technical specifications (HQMF, CQL, and QDM) specifications release format.
- Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid (and this must be demonstrated empirically).
- If a developer is testing an eCQM using any type of normalized EHR clinical data (e.g. from multiple EHR sources), NQF requires, at a minimum, supporting information of what schemas are included in the normalized data set and how they are calculated by the measure logic (i.e., what fields have been normalized and how, including any considerations of how this may affect the measure).
- **As of August 2019**, validity testing at the data element level will be required for all eCQMs. However, as with other measures, testing at the level of the performance measure score also is encouraged if data can be obtained from enough measured entities. If data element testing is not possible, justification is required and must be accepted by the Standing Committee.
    - If the testing is focused on validating the accuracy of the electronic data, analyze agreement between the electronic data obtained using the eCQM specifications and those obtained through abstraction of the entire electronic record (not just the fields used to obtain the electronic data), using statistical analyses such as sensitivity and specificity, positive predictive value, and negative predictive value. The guidance on measure testing allows this type of validity testing to also satisfy the requirement for reliability testing (see Algorithms 2 and 3).
    - Note that testing at the level of data elements requires that all critical data elements be tested (not just agreement of one final overall computation for all patients). At a minimum, the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately.
    - Use of a simulated data set (e.g. BONNIE) is no longer accepted for testing validity of data elements and is best suited for checking that the measure specifications and logic are working as intended and that value sets are included in the VSAC.
    - NQF's guidance has some flexibility; therefore, measure developers should consult with NQF staff if they think they have another reasonable approach to testing reliability and validity.

- The general guidance on samples for testing any measure also is relevant for eCQMs:
    - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
    - The sample should represent the variety of entities whose performance will be measured. The 2010 Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
    - The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
    - When possible, units of measurement and patients within units should be randomly selected.

- The following subcriteria under Scientific Acceptability of Measure Properties also apply to eCQMs.
    - Exclusion analysis (2b2). If exclusions (or exceptions) are not based on the clinical evidence, analyses should identify the overall frequency of occurrence of the exclusions as well as variability across the measured entities to demonstrate the need to specify exclusions.
    - Risk adjustment (2b3). Outcome and resource use measures require testing of the risk adjustment approach.
    - Differences in performance (2b4). This criterion is about using the measure as specified to distinguish differences in performance across the entities that are being measured. The performance measure scores should be computed for all accountable entities for which eCQM data are available (not just those on which reliability/validity testing was conducted) and then analyzed to identify differences in performance.
    - Because eCQMs are submitted as separate measures, even if the same or similar measures exist, comparability of performance measure scores if specified for multiple data sources (2b5) does not apply.
    - Analysis of missing data (2b6). Approved recommendations from the 2012 projects on eCQM feasibility assessment, composites, and patient-reported outcomes call for an assessment of missing data or nonresponses.

eCQM Approval for Trial Use

Developers have indicated that it can be challenging to test eCQMs to the extent necessary to meet NQF endorsement criteria—at least until they have been more widely implemented. At the same time, there is interest in developing eCQMs for use in federal programs and obtaining NQF endorsement for those eCQMs. NQF endorsement may provide the impetus to implement measures; however, if a submitted measure with very limited testing does not meet NQF endorsement criteria, it could be prematurely abandoned.

The ***Trial Use Program*** is specifically designed for eCQMs that are ready for implementation but cannot yet be adequately tested to meet NQF endorsement criteria. The program seeks to identify and support eCQMs that address important areas of performance measurement and quality improvement. To be included in the program, eCQMs must be assessed as technically acceptable for implementation and developers must have a plan to conduct more robust reliability and validity testing that takes advantage of clinical data in EHRs.

Candidacy for the Trial Use program is similar to eCQM endorsement candidacy and is reviewed by the standing committee based on the standard NQF evaluation criteria. Approved for Trial Use carries no endorsement label but may be considered a pathway for measures to prepare for endorsement. eCQMs that are Approved for Trial Use are indexed in QPS and are indicated as part of the program. See **Table 11** for comparison of endorsement and approval for trial use.

Candidates for the eCQM Trial Use Program are initially screened by the NQF eCQM and Maintenance teams prior to standing committee review and consideration. Initial screening includes:

- Must meet all criteria under Importance to Measure and Report (clinical evidence and opportunity for improvement/ performance gap)
- Completion of the eCQMs feasibility assessment, including NQF Feasibility Scorecard and simulated data set results (from BONNIE or another source)
- Plan for future use and discussion of how these measures will be useful for accountability and improvement
- Identification of related and competing measures with a plan for harmonization or justification for developing a competing measure

Maintenance of Trial eCQMs

Approved for Trial Use designation expires three years after initial committee approval date (if the eCQM is not submitted for endorsement prior to that time).

- There is no expectation that every trial use measure will be submitted for endorsement consideration – some may fail during testing.
- If submitted for endorsement three or more years after the Approval for Trial Use date, the measure must be submitted and evaluated on all criteria, similar to any measure being submitted for initial endorsement consideration.

eCQMs approved for Trial Use may be submitted for endorsement prior to the three-year expiration. The developer can select from the following options for evaluation and endorsement:

- **Option 1**: Submit and evaluate only *Scientific Acceptability of Measure Properties*, including the final eCQM specifications and all testing. If endorsed, endorsement date will assume the Approved for Trial Use date. Endorsement maintenance will be scheduled on the regular three-year cycle and the measure will be subject to evaluation on all criteria.
- **Option 2**: Submit and evaluate on all criteria. If endorsed, a new endorsement date will be identified and endorsement maintenance will be scheduled from the new endorsement date, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria.

Table 11. Endorsement versus eCQM Trial Use Approval

	Endorsement	eCQMs Trial Use Approval
Meaning	The eCQM has been judged to meet all NQF evaluation criteria and is suitable for use in accountability applications as well as performance improvement.	The eCQM has been judged to meet the criteria that indicate its readiness for implementation in real-world settings in order to generate the data required to assess reliability and validity.
Measure Evaluation	Reliability and validity testing results are required upon submission. All criteria are voted on by the Committee. Measure information forms for all measures under review for endorsement are made available on the project webpage.	Reliability and validity testing results are not needed for submission. All other criteria are voted on by the Committee. Measure information forms for all eCQMs under review for Trial Use Approval are made available on the project webpage.
Public and Member Comment	Same process. Comments may be submitted on measures recommended and not recommended for endorsement. NQF members may express support ("Support" or "Do Not Support") for each measure.	Same process. Comments may be submitted on eCQMs recommended and not recommended for eCQM Trial Use Approval. NQF members may express support ("Support" or "Do Not Support") for each eCQM.
CSAC	Same process.	Same process.
Information in QPS	Specs for endorsed measures are available.	Specs for eCQMs recommended for Trial Use Approval are available.
Status	When due for maintenance review, the measure will be evaluated through the multistakeholder process.	Trial Use Approval designation expires 3 years after initial approval. When submitted for endorsement, the measure will require testing results and will be evaluated through the multistakeholder process. There are 2 options if submitted for endorsement prior to 3 year expiration: Option 1: Submit and evaluate only *Scientific Acceptability of Measure Properties*, including the final eCQM specifications and all testing. If endorsed, endorsement date will assume the Approved for Trial Use date. Endorsement maintenance will be scheduled on the regular three-year cycle and the measure will be subject to evaluation on all criteria. Option 2: Submit and evaluate on all criteria. If endorsed, a new endorsement date will be identified and endorsement maintenance will be scheduled from the new endorsement date, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria.

Guidance for Considering Adjustment for Social Risk Factors

Guidance for Measure Developers

Background Information on the SDS Trial Period

- In late 2014, NQF's Board of Directors approved a 2-year trial period for risk adjustment for social risk factors prior to a permanent change in NQF policy.
- During the trial period, the NQF policy that restricted use of social risk factors in risk-adjustment approaches was suspended, and NQF implemented several of the Risk Adjustment Expert Panel's recommendations.
- The initial SDS Trial concluded in Spring 2017. After review of the findings of the trial, NQF's Board of Directors agreed to allow, for the present, use of social risk factors in risk-adjustment approaches. A second Social Risk Trial began in 2017 and will run until 2021. As in the first trial, measure developers **are required** to provide a conceptual rationale for how a social risk factor affects an outcome of interest. If a conceptual relationship exists, developers **should conduct empirical analyses** to examine the relationship between the social risk factor and the outcome of interest.

Instructions for providing required information on inclusion of social risk factors in risk adjustment

NOTE: These instructions ***are applicable to all*** health outcome measures, instrument-based measures (including patient-reported outcome based performance measures (PRO-PMs)), and intermediate outcome measures, and are potentially applicable to some process measures.

- Enter patient-level social risk variables that were available and analyzed during measure development in Section **1.8** of the Measure Testing Attachment. These variables could include:
    - Patient-reported data (e.g., income, education, language)
    - Proxy variables when social risk data are not collected from each patient (e.g., based on patient address and use of census tract data to assign individual patients to a category of income, education, etc.) and conceptual rationale for use
    - Patient community characteristics (e.g., crime rate, percent vacant housing, smoking rate, level of uninsurance) assigned to individual patients for the specific community where they live (not in the community in which the healthcare unit is located) [NOTE that these do not have to be a proxy for patient-level data]
- If you ARE risk-adjusting your measure, in addition to the conceptual/clinical and statistical methods and criteria used to select patient risk factors, describe the conceptual description (logical rationale or theory informed by literature and content experts) of the pathway between the patient social risk factors, patient clinical factors, quality of care, and outcome in Section **2b3.3a** of the Measure Testing Attachment. In Section **2b3.3b** of the Measure Testing Attachment, indicate how the conceptual model was developed.
- If you are NOT risk-adjusting your measure, include discussion of, and data for, social risk factors as part of the rationale and analysis included in Section **2b3.2** of the Measure Testing Attachment.
- Enter the analyses and interpretation resulting in the decision to include or not include social risk factors in Section **2b3.4b** of the Measure Testing Attachment. This analysis could include:
    - Variation in prevalence of the factor across measured entities
    - Empirical association with the outcome (univariate)
    - Contribution of unique variation in the outcome in a multivariable model

    - Assessment of between-unit effects vs. within-unit effects to evaluate potential clustering of disadvantaged patients in lower quality units
    - Impact of adjusting for social risk (or not) on providers at high or low extremes of social risk
- Enter reliability and validity testing for the measure as specified in Sections **2a2** and **2b1** of the Measure Testing Attachment.
    - If changing from a risk adjustment model that did not include social risk factors to one that does include social risk factors, then updated reliability and validity testing is required and must be entered into section 2a2 and 2b2 of the Measure Testing Attachment.
- Enter a comparison of performance scores with and without social risk factors in the risk adjustment model in Section **2b5** of the Measure Testing Attachment.
    - In Section **2b5.1**, enter the method of testing conducted to compare performance scores with and without social risk factors in the risk adjustment model for the same entities. Describe the steps and the statistical approach used.
    - In Section **2b5.2**, enter the statistical results from testing the differences in the performance scores with and without social risk factors in the risk adjustment model. (e.g., correlation, rank order)
    - In Section **2b5.3**, provide an interpretation of your results in terms of the differences in performance scores with and without social risk factors in the risk adjustment model for the same entities. What do the results mean, and what are the norms for the test conducted?
    - NOTE: If the measure has more than one set of specifications/instructions (e.g., one for medical record abstraction and one for claims data), then section 2b6 must also be used to demonstrate comparability of the performance scores.
- If a performance measure includes social risk variables in its risk adjustment model, the measure developer must provide the information required to stratify a clinically-adjusted only version of the measure results for those social risk variables in Section **S.11** in the Measure Submission Form. This information should *include* the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate.
- Enter the details of the final statistical risk model and variables in Section **2b3.1.1** of the Measure Testing Attachment.

Guidance on Evaluating Instrument-Based Measures, Including Patient-Reported Outcome Performance Measures (PRO-PMs)

See NQF report **Patient-Reported Outcomes in Performance Measurement (December 2012).**

Table 12. Distinctions among PRO, PROM, and PRO-PM: Two Examples

Definition	Patients with Clinical Depression	Persons with Intellectual or Developmental Disabilities
Patient-reported outcome (PRO): The concept of any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. PRO domains encompass: • health-related quality of life (including functional status); • symptom and symptom burden; • experience with care; and • health behaviors.	Symptom: depression	Functional Status-Role: employment
PRO measure (PROM): Instrument, scale, or single-item measure used to assess the PRO concept as perceived by the patient, obtained by directly asking the patient to self-report (e.g., PHQ-9).	PHQ-9©, a standardized *tool* to assess depression	Single-item measure on National Core Indicators Consumer Survey: *Do you have a job in the community?*
PRO-based performance measure (PRO-PM): A performance measure that is based on PROM data aggregated for an accountable healthcare entity (e.g., percentage of patients in an accountable care organization whose depression score improved as measured by the PHQ-9).	Percentage of patients with diagnosis of major depression or dysthymia and initial PHQ-9 score >9 with a follow-up PHQ-9 score <5 at 6 months (NQF #0711)	The proportion of people with intellectual or developmental disabilities who have a job in the community

Table 13. NQF Endorsement Criteria and their Application to Instrument-Based Measures

Abbreviated NQF Endorsement Criteria	Considerations for Evaluating instrument-based measures that are relevant to other performance measures	Unique Considerations for Evaluating instrument-based measures
1. Importance to Measure and Report a. Evidence: Health outcome OR evidence-based intermediate outcome, process, or structure of care b. Performance gap c. Composite	• PRO-PMs should have the same evidence requirement as health outcomes, i.e., empirical data demonstrates the relationship of the health outcome to processes or structures of care. • Process or structure measures derived from data collected via instrument have the same evidence requirements as other structure or process measures (i.e., a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence linking the measured structure or process to a desired outcome). • Exceptions to the evidence requirement for performance measures focused solely on administering a particular instrument should be addressed the same way as for other measures based solely on conducting an assessment (e.g., order lab test, check BP).	• Patients/persons must be involved in identifying structures, processes, or outcomes for performance measurement (person-centered; meaningful).

Abbreviated NQF Endorsement Criteria	Considerations for Evaluating instrument-based measures that are relevant to other performance measures	Unique Considerations for Evaluating instrument-based measures
2. Scientific Acceptability of Measure Properties a. Reliability 1. Precise specifications 2. Reliability testing (data elements or performance measure score) b. Validity 1. Validity testing (data elements or performance measure score) 2. Exclusions 3. Risk adjustment 4. Identify differences in performance 5. Comparability of multiple sets of specifications 6. Missing data/non-response	• Data collection instruments (tools) should be identified (e.g., specific instrument, scale, or single item). • If multiple data sources (i.e., instruments, methods, modes, languages) are used, then comparability or equivalency of performance measure scores should be demonstrated.	• Specifications should include standard methods, modes, languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; how missing data are handled; and calculation of response rates to be reported with the performance measure results. • Reliability and validity should be demonstrated for <u>both</u> the data (instrument) and the performance measure score. • Differences in individuals' responses related to instruments or methods, modes, and languages of administration need to be analyzed and potentially included in risk adjustment. • Response rates can affect validity and should be addressed in testing.
3. Feasibility a. Data generated and used in care delivery b. Electronic data c. Data collection strategy can be implemented	• The burdens of data collection, including those related to use of proprietary instruments, are minimized and do not outweigh the benefit of performance measurement.	• The burden to respondents (people providing the data) should be minimized (e.g., availability and accessibility enhanced by multiple languages, methods, modes). • Infrastructure to collect instrument-level data and integrate into workflow and EHRs, as appropriate.
4. Usability and Use 4a. Use 4a1. Accountability and transparency 4a2. Feedback by those being measured/others 4b. Usability 4b1. Improvement 4b2. Benefits outweigh unintended negative consequences	• Adequate demonstration of the criteria supports usability and ultimately the use of an instrument-based measure for accountability and performance improvement.	
5. Comparison to Related or Competing Measures **5a.** Harmonization of related measures **5b.** Competing measures	• Apply to the instrument-based performance measures	• Performance measure specified to use different instruments will be considered competing measures

Guidance on Evaluating Composite Performance Measures

Definition

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.

Identification of Composite Performance Measures for Purposes of NQF Measure Submission, Evaluation, and Endorsement

The listing below 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. See **Table 14** for details on the evaluation criteria for composite measures.

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, including
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient);

The following **will not be** considered composite performance measures 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 to 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.
- 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).

Table 14. NQF Measure Evaluation Criteria and Guidance for Evaluating Composite Performance Measures

Abbreviated NQF Endorsement Criteria	Guidance for Composite Performance Measures
1. Importance to Measure and Report a. Evidence: Health outcome OR evidence-based intermediate outcome, process, or structure of care b. Performance gap c. For composite performance measures, the following must be explicitly articulated and logical: 1. 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 2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and 3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.	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). 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 included in the composite is required (e.g., increase reliability of the composite, clinical evidence). 1c. Must also be met for a composite performance measure to meet the must-pass criterion of Importance to Measure and Report. If the developer provides a conceptual justification as to why an "any-or-none" 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.
2. Scientific Acceptability of Measure Properties a. Reliability 1. Precise specifications 2. Reliability testing (data elements or performance measure score) b. Validity 1. Validity testing (data elements or performance measure score) 2. Exclusions 3. Risk adjustment 4. Identify differences in performance 5. Comparability of multiple sets of specifications 6. Missing data/nonresponse 2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that: 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 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).	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 or differential weighting when combined into the composite). **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, interunit reliability, and intraclass correlation coefficient. 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. **2b1.** 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

Abbreviated NQF Endorsement Criteria	Guidance for Composite Performance Measures
	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. **2b2.** Applies to the component measures and composite performance measures. **2b3.** Applies to outcome component measures (unless NQF-endorsed). **2b4.** Applies to composite performance measures. **2b5.** Applies to component measures. **2b6.** Analyses of 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. **2c**. Must also be met for a composite performance measure to meet the must-pass criterion of Scientific Acceptability of Measure Properties. 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 criterion of Scientific Acceptability of Measure Properties. **Examples** of analyses: 1. ***If components are correlated*** – analyses based on shared variance (e.g., factor analysis, Cronbach's alpha, item-total correlation, mean inter-item correlation). 1. ***If components are not correlated*** – 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,[10] or clinical justification (e.g., correlation of the individual component measures to a common outcome measure). 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.

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

Abbreviated NQF Endorsement Criteria	Guidance for Composite Performance Measures
3. Feasibility a. Data generated and used in care delivery b. Electronic data c. Data collection strategy can be implemented	**3a, 3b, 3c.** Apply to composite performance measures as a whole, taking into account all component measures.
4. Usability and Use 4a. Use 4a1. Accountability and transparency 4a2. Feedback on measure 4b. Usability 4b1. Improvement 4b2. Benefits outweigh unintended negative consequences	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). **4a1.** 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. **4a2.** Applies to composite performance measures (may also apply to component measures). **4b1.** Applies to composite performance measures. **4b2.** 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.
5. Comparison to Related or Competing Measures **5a.** Harmonization of related measures **5b.** Competing measures	**5a and 5b.** Apply to composite performance measures as a whole as well as the component measures.

Guidance for Evaluating Evidence for Measures of Appropriate Use

Measures for appropriate use of procedures and medical technologies are becoming more common and reflect multistakeholder interest in assessing appropriate use of healthcare services. Current NQF criteria and guidance regarding appropriate use measures indicate the following:

- NQF measure evaluation criteria state that evidence for measures that focus on inappropriate use should include "a systematic assessment and grading of the quality, quantity, and consistency of the body of evidence that the measured process *does not* lead to a desired health outcome." Thus, the evidence for appropriate/inappropriate use measures should primarily focus on the *lack of effectiveness or benefit* of the test or procedure to patients. Patient safety considerations such as unnecessary exposure to radiation or anesthesia, or complications from inappropriate tests or procedures, may contribute to the risk-benefit evidence.
- Cost and resource use are **not** the focus of appropriate use measures. The cost and resource use implications of appropriate use measures are no different than for other measures; for example, improvement in adverse outcomes after surgery will likely reduce costs; and improved use of screening tests will increase costs, but this is not a consideration for evaluating the measures.
- Appropriate use measures are not efficiency measures as currently defined by NQF (i.e., efficiency measures per the current NQF definition have both a quality component and a cost component in the measure construct).

Development of Appropriate Use Method

In the 1980s, RAND/UCLA developed a methodology to determine "appropriateness" of healthcare tests, procedures, and processes. This method has been used worldwide in a variety of medical applications and forms the basis of many appropriate use measures (AUM) submitted to NQF. The RAND/UCLA Appropriateness Method User's Manual (2001) defines

> An appropriate procedure as one in which "the expected health benefit (e.g., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (e.g., mortality, morbidity, anxiety, pain, time lost from work) by a sufficiently wide margin that the procedure is worth doing, exclusive of cost…."
>
> The rationale behind the method is that randomized clinical trials—the "gold standard" for evidence-based medicine—often either are not available or cannot provide evidence at a level of detail sufficient to apply to the wide range of patients seen in everyday clinical practice. Although robust scientific evidence about the benefits of many procedures is lacking, physicians must nonetheless make decisions every day about when to apply them. Consequently, the RAND/UCLA researchers believed a method was needed that would combine the best available scientific evidence with the collective judgment of experts to yield a statement regarding the appropriateness of performing a procedure at the level of patient-specific symptoms, medical history and test results."

Various specialty societies such as the American College of Radiology and the American College of Cardiology Foundation/American Heart Association have used the RAND/UCLA methodology to develop appropriate use criteria for imaging and cardiovascular technology. The American Academy of Orthopedic Surgeons and the American Academy of Dermatology have also established appropriate use criteria for aspects of their specialty. These specialty society guidelines are intended to guide clinicians in the appropriate use of various tests and procedures.

Clinical Practice Guidelines and Appropriate Use Criteria

The appropriate use criteria are guidelines for clinical practice. The method for developing appropriate use criteria is very similar to the method used to develop traditional clinical practice guidelines (CPGs). Table 15 presents a side-by-side comparison of the methods for developing CPGs and Appropriate Use Criteria (AUC). Development of both types of guidelines is based on a review of the evidence.

Table 15. Comparison of Development of CPGs and AUCs

Clinical Practice Guidelines	Appropriate Use Criteria
Generally disease- or condition-based	Generally procedure- or test-based
Methodology: Institute of Medicine "Clinical Practice Guidelines We Can Trust" "The processes by which a CPG is developed and funded should be detailed explicitly and publicly accessible."	Methodology: RAND/UCLA Appropriateness Method (RAM)
Evidence review: CPG developers should use systematic reviews that meet standards set by the IOM's Committee on Standards for Systematic Reviews of Comparative Effectiveness Research: • A summary of relevant available evidence (and evidentiary gaps), description of the quality (including applicability), quantity (including completeness), and consistency of the aggregate available evidence. • A clear description of potential benefits and harms. • A rating of the level of confidence in (certainty regarding) the evidence underpinning the recommendation.	Evidence review: • Fundamental to any appropriateness study is a critical review of the literature summarizing the scientific evidence available on the procedure under review. Literature reviews for appropriateness studies are typically less strict in their inclusion criteria, as the objective is to produce a synthesis of all the information available on a particular topic; where evidence from controlled trials is lacking, they may well include lower-quality evidence from, for example, cohort studies or case series. • Where possible, "evidence tables" summarizing the data from multiple studies should be included in the literature review.
Guideline development group (GDG) composition: • The GDG should be multidisciplinary and balanced, comprising a variety of methodological experts and clinicians, and populations expected to be affected by the CPG. • Whenever possible GDG members should not have conflicts of interest (COI). • Funders should have no role in CPG development.	Expert panel: • Most users of the RAND/UCLA method recommend using multidisciplinary panels to better reflect the variety of specialties that are actually involved in patient treatment decisions. • The RAM is a modified Delphi method that, unlike the original Delphi, provides panelists with the opportunity to discuss their judgments between the rating rounds.

Clinical Practice Guidelines	Appropriate Use Criteria
Guideline Recommendations: • Recommendations should include an explanation of the reasoning underlying the recommendation. • A rating of the strength of the recommendation in light of the evidence. • A description and explanation of any differences of opinion regarding the recommendation. • Recommendations should be articulated in a standardized form detailing precisely what the recommended action is and under what circumstances it should be performed. • Strong recommendations should be worded so that compliance with the recommendation(s) can be evaluated. • The CPG publication date, date of pertinent systematic evidence review, and proposed date for future CPG review should be documented in the CPG.	RAND/UCLA Appropriateness Method (RAM): • A list of the hypothetical clinical scenarios or "indications" to be rated by the panel is developed. The purpose of the list of indications is to classify patients in terms of the clinical variables physicians take into account in deciding whether to recommend a particular procedure. • Panelists are asked to rate the appropriateness of each indication using their own best clinical judgment (rather than their perceptions of what other experts might say) and considering an average patient presenting to an average physician who performs the procedure in an average hospital (or other care-providing facility). *They are specifically instructed not to consider cost implications in making their judgments*. Although cost considerations are an important factor in deciding whether a procedure or treatment should ultimately be made available to patients, the RAM focuses on the initial question of whether it is effective. • In the RAM, a procedure is classified as "appropriate," "uncertain," or "inappropriate" for a particular patient scenario ("indication") in accordance with 1) the *median* panel rating and 2) some measure of the dispersion of panel ratings, which is taken as an indicator of the level of agreement with which the ratings were made. [This is not a consensus process.]

NQF's Evaluation Criteria for Evidence

NQF's guidance for evidence for measures in general, and specifically those based on clinical practice guidelines, applies to measures based on appropriateness criteria as well. As noted in Table 15 above, both CPGs and appropriateness methodologies require systematic reviews of the evidence generated from a thorough literature search.

Measure Submission

Measure submitters should provide the information on evidence that was provided to the expert panel that developed the appropriate use criteria, along with any updated evidence published since the AUC was developed. The measure submission should include:

- a summary (not a list of references) of the evidence in the submission evidence attachment that describes the quantity, quality, and consistency of the body of evidence (not selected references) and an assessment of the benefits versus harms; and

- a link to (or an attached appendix that contains) the complete evidence report, with evidence tables, if available.

Committee Evaluation

Committees should review the information provided and evaluate the evidence presented according to the Algorithm 1.

- It is unlikely that a systematic review will have been performed to establish a lack of benefit for an intervention. Begin at Box 7 – empiric evidence submitted without systematic review and grading of the evidence.
- If a complete literature review is summarized (rather than selected studies -Box 8) then the Committee should decide whether the submitted evidence indicates a **high certainty** and that benefits clearly outweigh undesirable effects (Box 9). If yes, then rate as moderate.
- If there is no empiric evidence, skip Box 10 and go to Box 11. The Committee should agree that the AUC method is a systematic assessment of expert opinion that the benefits of what is being measured outweigh the potential harms (Box 11). If the Committee agrees that it is acceptable (or beneficial) to hold providers accountable for performance in the absence of empiric evidence (Box 12), then rate as "insufficient evidence with exception."

Guidance for Population Health and Access Measures[11]

Background

Access to care is essential, particularly for our currently fragmented healthcare system, which generally delivers episodes of face-to-face treatment with minimal communication between encounters. While people agree that access to healthcare is necessary, there are several definitions and interpretations of access to care, creating confusion and frustration for all. Moreover, measuring access is further confounded by interpreting what is meaningful access, what care actually was delivered, the timeliness of care, and the impact of access on intermediate outcomes or outcomes. Measuring the quality of services differs from measuring the access to services of different quality levels.

Access often is associated with the availability of resources and frequently depends on financing. Penchansky and Thomas describe access as a "set of dimensions that characterize the fit between the patient and the healthcare system," including geographical, temporal, financial, cultural, and digital access.[12]

Traditional access concepts, and hence measurement points, focus on in-person experiences between the patient and provider; an array of historical frameworks present models of access that are useful for thinking

[11] This guidance was developed initially as part of the 2016 off-cycle work of the Health and Well-Being Standing Committee.

[12] Khan AA, Bhardwaj SM. Access to health care: a conceptual framework and its relevance to health care planning. *Eval Health Prof*. 1994;17(1):60-76.

about measuring access.[13,14,15,16] However, there are opportunities beyond these paradigms. One potential option is to improve digital access between the patient and provider.[17] A shift in culture will be required to utilize this method, which could help diminish geographical, temporal, and cultural access problems faced by patients. For example, NQF's work on performance measures for rural providers noted that telehealth and telemedicine allow greater access to care; thus, permitting telehealth and telemedicine to "count" as successfully meeting clinical measures serves quality improvement, as well as access.[18]

Access to healthcare also can be improved beyond the doctor's office or hospital by providing wellness and health promotion at work sites, which is where many individuals spend the majority of their time, or through health system changes and a focus on population health as the measurement leverage point.[19] Measuring access to healthcare also can be leveraged by examining modifiable financial (e.g., underinsurance), structural (e.g., transportation, waiting times, access to primary care or safety net institutions), and cognitive barriers (e.g., health literacy, interpreter services) that apply broadly, but are especially important to reducing disparities.[20]

NQF works to help improve access to care by both seeking to endorse performance measures that can help identify key areas to measure access and identifying gaps in access-to-care measures. During the Health and Well-Being Phase 2 project, the Standing Committee noted the measurement focus and specifications of measures #1516, #1392, #2689, and #2695[21] do not capture whether specific care processes occur during a patient encounter, rather only confirm the visit[22]—even though the developer(s) explicitly stated that these measures are intended to assess access to care. As an example, the two well-child visit measures assess only that visits occurred and not whether the child received the age-appropriate vaccinations, hearing, or vision tests. Other measures were focused more globally, e.g., hospitalization for dehydration, and were asserted to reflect access to and coordination of a community's ambulatory services.

The purpose of this document is to provide guidance to developers and NQF Committees on access-to-care

---

[13] Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? *J Health Soc Behav*. 1995;36(1):1-10.

[14] Flores G, Vega JR. Barriers to health care access for Latino children: a review. *Fam Med*. 1998;(3):196-205.

[15] Fitzpatrick AL, Powe NR, Cooper LS, et al. Barriers to health care access among the elderly and who perceives them. *Am J Public Health*. 2004;94(10):1788-1794

[16] DeVoe JE, Baez A, Angier H, et al. Insurance+access not equal to health care: typology of barriers to health care access for low-income families. *Ann Fam Med*. 2007;5(6):511-518.

[17] Fortney JC, Burgess JF Jr, Bosworth HB, et al. Re-conceptualization of access for 21st century healthcare. *J Gen Intern Med*. 2011;26 (Suppl 2):S639-S647.

[18] National Quality Forum (NQF). Performance Measurement for Rural Low-Volume Providers. Final Report. Washington, DC: NQF; 2015. Available at http://www.qualityforum.org/Publications/2015/09/Rural_Health_Final_Report.aspx. Last accessed July 2016.

[19] Stoto M. Population Health Measurement: Applying Performance Measurement Concepts in Population Health Settings. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2015;2(4):Article 6.

[20] Carillo JE, Carillo VA, Perez HR, et al. Defining and targeting health care access barriers. *J Health Care Poor Underserved*. 2011;22:562-575.

[21] NQF 1516 Well-Child Visits in the Third, Fourth, Fifth, and Sixth Years of Life; NQF 1392

Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)

[22] National Quality Forum (NQF). Health and Well-Being Phase 2. Final Report. Washington, DC: NQF; 2015. Available at http://www.qualityforum.org/Publications/2015/11/Health_and_Well-Being_Phase_2_Final_Report.aspx. Last accessed May 2016.

measure development and the NQF evaluation of such measures. **Table 16** also includes a few examples of measures and concepts that NQF developers and others identify as reflecting access to care (ambulatory care sensitive emergency department visits for dental caries in children), some of which are closer to the "access event" and others further away—which likely involve other factors (e.g., dehydration admissions) in addition to access.

Table 16. Examples of Existing Access Measures and Concepts

Subject/Concept	Measure Title	Steward
Dental Care Visits	1) Ambulatory Care Sensitive Emergency Department Visits in Dental Caries in Children[23] 2) Follow-Up after Emergency Department Visit by Children for Dental Caries[24]	American Dental Association/Dental Quality Alliance
Well-Child Visits	1) Well-Child Visits in the Third, Fourth, Fifth, and Sixth Years of Life[25] 2) Well-Child Visits in the First 15 Months of Life[26]	National Committee for Quality Assurance
Care Coordination for Children with Complex Medical Needs	Family Experiences with Coordination of Care (FECC)-1 Has Care Coordinator[27]	Seattle Children's Research Institute
Prenatal and Postpartum Care	1) Timeliness of Prenatal Care 2) Postpartum Care[28]	National Committee for Quality Assurance
Dehydration Admissions	Dehydration Admission Rate (PQI 10)[29]	Agency for Healthcare Research and Quality
Patient Reporting of Access to Services, Cognitive Barriers	CAHPS Clinician & Group Surveys (CG-CAHPS)-Adult, Child[30]	Agency for Healthcare Research and Quality
HIV/AIDs	HIV Late Diagnoses[31]	Centers for Disease and Control and Prevention
Health Insurance Coverage	Percent of persons with health insurance	NHIS (national database)*

[23] NQF 2689: Ambulatory Care Sensitive Emergency Department Visits for Dental Caries in Children

[24] NQF 2695: Follow-Up after Emergency Department Visit by Children for Dental Caries

[25] NQF 1516: Well-Child Visits in the Third, Fourth, Fifth, and Sixth Years of Life

[26] NQF 1392: Well-Child Visits in the First 15 Months of Life

[27] NQF 2842: Family Experiences with Coordination of Care (FECC)-1 Has Care Coordinator

[28] NQF 1517: Prenatal and Postpartum Care

[29] NQF 0280: Dehydration Rate (PQI 10)

[30] NQF 0005: CAHPS Clinician & Group Surveys (CG-CAHPS)-Adult, Child

[31] NQF 1999: Late HIV Diagnoses

Subject/Concept	Measure Title	Steward
Unmet Need	Percent of families that experience difficulties or delays in obtaining health care or do not receive needed care for one or more family members	MEPS/MCBS (national database)*
Mental Health/Substance Abuse	Percent of adults with serious mental illness who received treatment	NHSDA (national database)*

*These measures are a part of AHRQ's preliminary measure set, National Healthcare Disparities Report, 2002: http://archive.ahrq.gov/research/findings/nhqrdr/nhdr02/premeasurea.html.

Overall, measures that focus directly on overcoming barriers (structural, financial, cognitive) to access and are closer to the "access event" are the most direct and desirable. Access measures also should advance one or more of the Institute of Medicine's six aims for healthcare—safe, effective, patient-centered, timely, efficient, and equitable.[32] Additionally, just as measurement for the pediatric population is generally under-represented, access measures for the pediatric population are encouraged (e.g., a pediatric corollary to the adult measure would be 'percent of children with serious mental illness who received treatment'). Finally, in considering access measures, framing an NQF access portfolio against the traditional categories of structure, process, and outcome measures[33] may provide guidance for future development activities, as well as identify gaps in access measures, generally, and in the portfolio, specifically (Table 17).

Table 17. Framing Future Access Measures

Structure	Process	Outcome
• Structures must be in place to access care (e.g., sufficient primary care, transportation, financing) • Access measures ideally address overcoming such structural barriers	• Processes must be in place to ensure access to care (e.g., follow-up) • Access measures ideally address the degree to which the process is adhered to	• Access is achieved (e.g., service is utilized) • Access measures ideally address appropriate and/or timely utilization

How to Develop, Review, and Evaluate Access Measures

Performance measures are traditionally evaluated against NQF's measure evaluation criteria, which are used to determine suitability of measures for use in both quality improvement efforts and for accountability purposes. The five major criteria[34] are:

[32] Institute of Medicine. *Crossing the Quality Chasm: A New health System for the 21st Century*. Washington, DC: National Academies Press; 2001.

[33] Donabedian, A. The quality of care: How can it be assessed? *JAMA.* 1998;121(11):1145-1150.

[34] More detail on these criteria can be found in the Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement Document.

1) Importance to measure and report – This criterion allows for a distinction between things that are important to do (or outcomes of importance) versus those processes, structures, or outcomes that rise to the level of importance required for a national performance measure. Importance has two key subcriteria: Evidence and Performance Gap. Evidence is the extent to which the specific measure focus is evidence-based and can drive significant gains in healthcare quality. Performance gap denotes that there is variation in performance among measured entities or that disparities (e.g., by race or ethnicity) exist even if a "macro-level" analysis appears to show that a measure is topped out.
2) Scientific acceptability of measure properties – This reflects NQF's view that performance measures must demonstrate sound measurement science—that is, they must be both reliable and valid.
3) Feasibility – The Feasibility criterion reflects the extent to which the data required to compute a measure are readily available and retrievable without undue burden, as well as the ease of implementation for performance measurement.
4) Usability and Use – NQF-endorsed measures are considered suitable for both accountability and quality improvement purposes, and the expectation is that endorsed measures not only will be used, but also ultimately will lead to improved patient outcomes.
5) Comparison to related or competing measures – Since there is an abundance of measures, this criterion requires a careful consideration of such similar measures, with the goal of endorsing only the best measures—or, if there is not a "best" measure, endorsing measures that are consistent to the extent possible.

Over time, NQF has evolved from its focus on traditional quality measures to include other measures of performance. For example, cost and resource use measures—the building blocks of measures of efficiency—complement quality measures. Access measures are similarly complementary and can address effectiveness, timeliness, efficiency, and/or disparities. Both types provide a better understanding of the overall performance of the healthcare system.

NQF has defined access as the "ability to obtain needed healthcare services in a timely manner including the perceptions and experiences of people regarding their ease of reaching health services or health facilities in terms of proximity, location, time, and ease of approach. Examples may include, but are not limited to, measures that address the timeliness of response or services, time until next available appointment, and availability of services within a community."[35] From this, a minimum scope of access measures could be inferred as addressing timeliness and availability. More broadly, NQF seeks access measures that address identified barriers, are reasonably close to the "access event," and will drive improvement in one or more of the six aims for healthcare quality and address basic principles of access to healthcare.[36]

Currently, the NQF portfolio lacks a robust set of measures related to access (defined by any means). Based on experience with other classes of measures, specific guidance on how NQF Committees should evaluate access measures can, in turn, provide clarity to developers on nuances of developing such measures and NQF's expectations for them.

Recognizing that the five core evaluation criteria are relevant, but require additional guidance for certain types

[35] National Quality Forum (NQF). *Glossary of terms*. Washington, DC: NQF; 2013.

[36] Institute of Medicine. Committee on Optimizing Scheduling in Health Care. Transforming Health Care Scheduling and Access. Washington, DC: National Academies Press; 2015.

of measures, NQF has provided additional guidance on composite, appropriate use, cost and resource use, population health, and patient-reported outcome measures. For population health measures, for example, NQF's guidance[37] document notes that the core criteria remain the same, but the language and direction are tailored. This document addresses guidance to developers and NQF Committees on access to care measures.

Table 18 sets forth NQF's general evaluation criteria (left column). To provide context for the types of changes made for NQF's different types of guidance, the middle column presents the guidance specifically approved for population health measures. The final column presents the guidance for access measures.

[37] The complete population health guidance document can be found at this link: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=70394 .

Table 18. NQF Criteria, Population Health Measure Criteria, and Access Measure Criteria

NOTE: There have been a few changes in criteria since this table was developed, and these changes have **NOT** been incorporated into this table. However, access measures must conform to current criteria. Contact NQF staff with any questions on how the revised criteria affect access measures.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
Conditions for Consideration Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**	**Conditions for Consideration** Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**	**Conditions for Consideration** Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**
A. The measure is in the public domain or a measure steward agreement is signed.	A. No change.	**A.** No change. (Here and hereafter, "no change" refers to no change from the general criteria.)
B. The measure owner/steward verifies that there is an identified responsible entity and a process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every three years.	**B.** The measure owner/steward verifies that there is an identified responsible entity or multistakeholder entities and a process to maintain and update the measure on a schedule that is commensurate with the rate of **population health innovation**, but at least every three years.	**B.** The measure owner/steward verifies there is an identified responsible entity or multi-stakeholder entities and a process to maintain and update the measure on a schedule that is commensurate with the rate of **policy- or structural-related access innovation**, but at least every three years.
C. The intended use of the measure includes <u>both</u> public reporting <u>and</u> quality improvement.	**C.** The intended use of the measure includes <u>both</u> public reporting <u>and</u> improvement in efforts to improve population health.	**C.** The intended use of the measure includes <u>both</u> public reporting <u>and</u> **improvement in efforts to improve access**.
D. The measure is fully specified and tested for reliability and validity.[1]	**D.** No change**.**	**D.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.	**E.** The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate. **Harmonization of related measures at the provider and population levels has been considered and addressed.**	**E.** No change.
F. The requested measure submission information is complete and responsive to the questions so that all the information needed to evaluate all criteria is provided.	**F.** No change.	**F.** No change.
Note	**Note**	**Note**
1. An eMeasure that has not been tested sufficiently to meet endorsement criteria may be eligible for Approval for Trial Use. Time-limited endorsement is no longer available.	1. No change.	1. No change.
Criteria for Evaluation	**Criteria for Evaluation**	**Criteria for Evaluation**
If all conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Usability*, and *Feasibility*. Not all acceptable measures will be equally strong on each set of criteria. The assessment of each criterion is a matter of degree. However, if a measure is not judged to have met minimum requirements for *Importance to Measure and Report* or *Scientific Acceptability of Measure Properties*, it cannot be recommended for endorsement and will not be evaluated against the remaining criteria.	No change.	No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
1. Impact, Opportunity, Evidence—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-impact aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to*** *meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria.*	**1. Impact, Opportunity, Evidence—Importance to Measure and Report:** Extent to which the specific measure focus is evidence-based, important to making significant gains in **population health, improving determinants of health and health outcomes of a population** for a high-impact aspect of **health** where there is variation in **(including geographic variation)** or overall less-than-optimal performance. ***Measures must be judged to*** *meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria.*	**1. Impact, Opportunity, Evidence—Importance to Measure and Report:** Extent to which the specific measure focus is evidence-based, important to making significant gains in **access to care leading to improved health outcomes** for a high-impact aspect of healthcare **or health** where there is variation in **(including geographic variation and structural, financial, and cognitive barriers)** or overall less-than-optimal performance. ***Measures must be judged to*** *meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria*

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
1a. Evidence to Support the Measure Focus The measure focus is a health outcome or is evidence-based, demonstrated as follows: • Health outcome[3]: a rationale supports the relationship of the health outcome to processes or structures of care. • Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome. • Patient 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 for the quality component as noted above. **AND**	**1a. Evidence to Support the Measure Focus** The measure focus is a health outcome or is evidence-based, demonstrated as follows: • Health outcome:[3] a rationale supports the relationship of **the health outcomes in the population to strategies to improve health.** • **Health determinant**, Intermediate outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome. • **Experience with care, services, or other health determinants**: evidence that the measured aspects of care are those valued by **people and populations** and for which the **respondent** is the best and/or only source of information OR that experience is correlated with desired outcomes. • Efficiency[6]: evidence for the quality component as noted above. **AND**	**1a. Evidence to Support the Measure Focus** The measure focus is evidence-based, demonstrated as follows: • Health outcome[3] and utilization: a rationale supports the relationship to overcoming an access barrier to achieving an improved health outcome • Intermediate outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to improved access to care and a desired health outcome. • **Experience with access to care or services**: evidence that the measured aspects are those valued by **people and populations** and for which the **respondent** is the best and/or only source of information OR that experience is correlated with desired outcomes. • Efficiency[6]: evidence for the quality **and access** component as noted above. **AND**

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
1b. Performance Gap Demonstration of quality problems and opportunity for improvement, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care). **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. **1c.** For composite performance measures, the quality construct rationale, and aggregation and weighting rules explicitly articulated and logical. **Notes** **2.** 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.	**1b. Performance Gap** Demonstration of opportunity for improvement in **health**, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in **health across providers (healthcare, public health, and other partners) and/or population groups, (including but not limited to disparities in care).** **Disparities** If **health disparities** 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** **No option for justification for lack of stratification.** **1c.** No change **Notes** **2.** No change	**1b. Performance Gap** Demonstration of opportunity for improvement in **access**, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in **access across providers (healthcare, public health, and other partners) and/or population groups, (including but not limited to disparities in care).** **Disparities** If disparities in **access to** 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** No change. **1c.** No change **Notes** **2.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
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.	**3.** Not applicable	**3.** Not applicable
4. 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.	**4. Population health determinants typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with stakeholder input) → provide intervention → evaluate impact on population health status. If the measure focus is one step in such a multistep process, the steps with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.**	**4. Access typically includes several leverage points: access to payment coverage; covered services; access to (timely) services; receipt of services; quality of service received → improved outcome. If the measure focus is less proximal to the receipt of services and quality, the step with the strongest evidence for the link to improved access should be selected as the focus of measurement. In addition to decreased care, key leverage points for which access measures can be represented are measures of late presentation of disease and lack of/decreased prevention.**
5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.	**5.** No change.	**5.** No change.
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).	**6.** No change.	**6.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
2. Reliability and Validity—Scientific Acceptability of Measure Properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to*** *meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.*	**2. Reliability and Validity—Scientific Acceptability of Measure Properties:** Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to*** *meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.*	**2. Reliability and Validity—Scientific Acceptability of Measure Properties:** Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to*** *meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.*
2a. Reliability	**2a. Reliability**	**2a. Reliability**
2a1. The measure is well defined and precisely specified[7] so it can be implemented consistently within and across organizations and allow for comparability. EHR measure specifications are based on the quality data model (QDM).[8]	**2a1.** The measure is well defined and precisely specified[7] so it can be implemented consistently within and across organizations, **multistakeholder groups, populations, or entities with shared accountability for health** and allow for comparability.	**2a1.** No change.[7,8]
2a2. Reliability testing[9] 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.	**2a2.** No change.[9]	**2a2.** No change.[9]
2b. Validity	**2b. Validity.**	**2b. Validity**
2b1. The measure specifications[7] 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.	**2b1.** No change.	**2b1.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
2b2. Validity testing[10] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.	**2b2.** Validity testing[10] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the effect of interventions to **improve population health**, adequately identifying differences in effectiveness.	**2b2.** Validity testing[10] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in **access**.
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[11];	**2b3.** Exclusions are supported by the **evidence**; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion;[11]	**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[11];
AND	**AND**	**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).[12]	If **individual or subgroup** preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure **or variation**; in such cases, the measure must be specified so that the information about **individual or subgroup** preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).[12]	No change.[12]

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
2b4. For outcome measures and other measures when indicated (e.g., resource use): • an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on 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[13,14]; and has demonstrated adequate discrimination and calibration **OR** rationale/data support no risk adjustment/ stratification.	**2b4.** For outcome measures and other measures when indicated (e.g., resource use): an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on factors that influence the measured outcome (but not factors related to **disparities in population health or health interventions**) and are present at start of care[13,14]; and has demonstrated adequate discrimination and calibration	**2b4.** For **access** measures, **access in general, risk adjustment is not appropriate[13,14] nor is level of attribution and analysis at the individual practitioner or group practice. Attribution of access measures is most appropriate at broader levels (e.g., community, health plan, population, ACOs).** **AND** **as appropriate, access measures should address disease acuity and appropriate triage (e.g., timeliness measures).**
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[15] differences in performance; **OR** there is evidence of overall less-than-optimal performance.	**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 meaningful[15] differences in performance **or variation across populations in improving health.** **OR** there is evidence of overall less-than-optimal performance **or significant variation across populations**.	**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[15] differences in performance **(i.e., access)**; **OR** there is evidence of overall less-than-optimal performance **(i.e., access)**.
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.	**2b6.** No change.	**2b6.** No change.
2b7. For eMeasures, composites, and PRO-PMs: missing data analysis	**2b7. No change**	**2b7. No change**
2c. For composite performance measures, empirical analyses support the composite construction approach	**2c. No change**	**2c. No change**

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
Notes	**Notes**	**Notes**
7. 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.	**7.** No change	**7.** No change.
8. EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.	**8.** N/A	**8.** EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
9. 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).	**9.** No change.	**9.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
10. 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.	**10.** No change.	**10.** No change.
11. 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.	**11.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, **multistakeholder groups, and populations** and sensitivity analyses with and without the exclusion.	**11.** No change.
12. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.	**12.** N/A	**12.** No change.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
13. Risk factors that influence outcomes should not be specified as exclusions.	**13.** Risk factors that influence outcomes should not be specified as exclusions.	**13.** Risk factors that influence **access** should not be specified as exclusions.
14. 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	**14.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in health determinants, such as race, socioeconomic status, or gender (e.g., poorer **health outcomes** of African American men with prostate cancer or inequalities in 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.	**14.** If incorporated, risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in **access to** 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.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent vs. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 vs. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.	**15.** With large enough sample sizes, small differences that are statistically significant may or may not be **practically meaningful**. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of **people** who received smoking cessation counseling (e.g., 74 percent vs. 75 percent) is meaningful; or whether a statistically significant difference of $25 in cost for an intervention (e.g., $5,000 vs. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers **or populations**.	**15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of **people** who received smoking cessation counseling (e.g., 74 percent vs. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 vs. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and find them useful for decisionmaking.	**3. Usability: Note: intended audiences can include community members and coalitions.**	**3. Usability:** No change.
3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for public reporting (e.g., focus group, cognitive testing) or rationale;	**3a.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for public reporting (e.g., focus group, cognitive testing) or rationale;	**3a**. No change.
AND	**AND**	**AND**
3b. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing quality improvement[16] (e.g., quality improvement initiatives) or rationale.	**3b.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing improvement[16] in **health determinants and/or population health** or rationale.	**3b.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing improvement[16] in **access** or rationale.
Note	**Note**	**Note**
16. 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.	**16.** An important outcome that may not have an identified improvement strategy still can be useful for informing **improvement in quality and/or population health** by identifying the need for and stimulating new approaches to improvement.	**16.** An important **measure** that may not have an identified improvement strategy still can be useful for informing **improved access** by identifying the need for and stimulating new approaches to improvement.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
4. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.	**4. Feasibility:** No change.	**4. Feasibility:** Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement.
4a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).	**4a.** No change for clinically oriented measures.	**4a.** No change for clinically oriented measures.
4b. 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.	**4b.** The required data elements are available in electronic health records, **personal health records, health information exchanges, population data bases**, or other electronic sources. If the required data are not available **in existing electronic sources**, a credible, near-term path to electronic collection is specified.	**4b.** The required data elements are available in electronic health records, **personal health records, health information exchanges, population health data bases,** or other electronic sources. If the required data are not available **in existing electronic sources**, a credible, near-term path to electronic collection is specified.
4c. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.	**4c**. Susceptibility to inaccuracies, errors, **inappropriate comparison across populations,** or unintended consequences and the ability to audit the data items to detect such problems are identified.	**4c.** No change
4d. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality,[17] etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).	**4d.** No change.[17]	**4d.** No change.[17]

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
Note	**Note**	**Note**
17. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.	**17.** All data collection must conform to laws regarding protected health information. **Confidentiality** is of particular concern with measures based on **individual** surveys and **for small populations**.	**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.
5. Comparison to Related or Competing Measures If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.	**5. Comparison to Related or Competing Measures** If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure. **Note: Complementary measures that address different improvement strategies are not considered competing measures.**	**5. Comparison to Related or Competing Measures** No change.
5a. The measure specifications are harmonized[18] with related measures; **OR** the differences in specifications are justified.	**5a.** No change. **OR** **5b.** No change.	**5a.** The measure specifications are harmonized[18] with related measures. **Complementary measures that address different strategies to improve access are not considered competing measures. For example, a Medicaid program measure of access to X service and a system measure of availability (or delivery) of same service would be complementary and not competing.** **OR** the differences in specifications are justified.

NQF Measure Evaluation Criteria	Population Health Measure Evaluation: Additional Guidance and Context	Access Measure Evaluation Criteria: Additional Guidance and Context
5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure); **OR** multiple measures are justified.	**5b.** No change.	**5b.** No change.
Note	**Note**	**Note**
18. Measure harmonization refers to the standardization of specifications for related measures with the same measure focus (e.g., *influenza immunization* of patients in hospitals or nursing homes); related measures with the same target population (e.g., eye exam and HbA1c for *patients with diabetes*); or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.	**18. Additional conceptualization needed for harmonization between clinical and population-level measures.**	**18. Additional conceptualization needed for harmonization among clinical, population, resource use, appropriate use, and access measures (i.e., is a broader NQF portfolio issue).**

Additional Guidance

As noted, performance measures are specified by developers and are evaluated against NQF's measure evaluation criteria. One important component of these specifications is the level of analysis—i.e., attribution to the accountable entity. As noted in the previous section, ideal access measures for the purpose of accountability should be viewed as representing a shared responsibility and be broadly attributed—i.e., not specified for the individual practitioner or even group. In particular, such health plan-, ACO-, or population-level measures should not be applied or implemented at non-endorsed levels of accountability *ex post facto*.

Inactive Endorsement with Reserve Status (November 2014)

Given the number of publicly reported measures with high levels of performance, reliable and valid measures of great importance may not retain NQF endorsement due to the lack of a performance gap. The purpose of an inactive endorsement with reserve status is to retain endorsement of reliable and valid quality performance measures that have overall high levels of performance with little variability so that performance could be monitored as necessary to ensure that performance does not decline. This status would apply only to highly credible, reliable, and valid measures that have high levels of performance due to incorporation into standardized patient care processes and quality improvement actions. The key issue for continued endorsement is the opportunity cost associated with continued measurement at high levels of performance—rather than focusing on areas with known gaps in care. Endorsement with reserve status retains these measures in the NQF portfolio for periodic monitoring, while also communicating to potential users that the measures no longer address high leverage areas for accountability purposes.

Measures with High Levels of Performance: Recommendations from the Evidence Task Force

The 2010 Evidence Task Force defined the term "topped out," meaning there are high levels of performance with little variation and, therefore, little room for further improvement. The Task Force did not recommend specific quantitative thresholds for identifying conformance with the subcriterion of opportunity for improvement (1b). Threshold values for opportunity for improvement would be difficult to standardize and depend on the size of the population at risk, the effectiveness of an intervention, and the consequences of the quality problem**.** For example, even modest variation would be sufficient justification for some highly effective, potentially life-saving treatments (e.g., certain vaccinations) that are critical to the public health.

The Task Force noted that, at the time of endorsement maintenance review, if measure performance data indicate overall high performance with little variation, then justification would be required for continued endorsement of the measure. The Consensus Standards Approval Committee (CSAC) added that the default action should be to remove endorsement unless there is a strong justification to continue endorsement. If a measure fails opportunity for improvement (1b), then it does not pass the threshold criterion, *Importance to Measure and Report,* and is therefore not suitable for endorsement.

Task Force Recommendations related to opportunity for improvement (1b) include the following:

- At the time of initial endorsement, evidence for opportunity for improvement generally will be based on research studies, or on epidemiologic or resource use data. However, at the time of review for endorsement maintenance, the primary interest is on the endorsed measure as specified, and the *evidence for opportunity for improvement should be based on data for the specific endorsed measure.*
- When assessing measure performance data for opportunity for improvement, the following factors should be considered:
    - number and representativeness of the entities included in the measure performance data;
    - data on disparities; and
    - size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.
- In exceptional situations, a strong justification for continued endorsement could be considered (e.g., ***evidence*** that overall performance will likely deteriorate if not monitored, or magnitude of potential harm if outcomes deteriorate while not being monitored).

Criteria for Assigning Inactive Endorsement with Reserve Status to Measures with High Levels of Performance

There is rarely evidence that performance will deteriorate if a measure is not monitored; therefore, some additional criteria are needed. The following criteria are to be used when there are concerns that performance will deteriorate, but no evidence. These criteria are intentionally rigorous so that the use of endorsement with reserve status is by exception.

- Evidence of little opportunity for improvement (1b), i.e., overall high level of performance with little variation. When assessing measure performance data for opportunity for improvement, the following factors should be considered:
    - distribution of performance scores;
    - number and representativeness of the entities included in the measure performance data;
    - data on disparities; and
    - size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.
- Evidence for measure focus (1a) – there should be strong direct evidence of a link to a desired health outcome; therefore, there would be detrimental consequence on patient health outcomes if performance eroded. Generally, measures more distant from the desired outcome have only indirect evidence of influence on the outcome and would not qualify for reserve endorsement status. For process and structure measures, the measure focus should be close to the desired outcome. Generally, measures of activities far from the desired outcome would not be eligible for reserve status.
- Reliability (2a) – high or moderate rating: Reliability has been demonstrated for the measure score.
- Validity (2b) – high or moderate: Validity has been demonstrated by empiric testing for the measure score (face validity not acceptable).
- The reason for high levels of performance is better performance, not an issue with measure construction/specifications (e.g., "documentation").
- Demonstrated usefulness for improving quality (e.g., data on trends of improvement and scope of patients and providers included).
- Demonstrated use of the measure (e.g., specific programs and scope of patients and providers included; would not grant inactive endorsement status for a measure that has not been used).
- If a measure is found to be "topped out", i.e., does not meet criteria for opportunity for improvement (1b), the measure will only be considered for inactive endorsement with reserve status. The measure must meet all other criteria as noted above; otherwise the measure should not be endorsed.

Maintenance of Inactive Endorsement with Reserve Status

Measures assigned inactive endorsement status will not be reviewed in the usual endorsement maintenance review cycle. During portfolio review, the Standing Committee will periodically review measures in reserve status for any change in evidence, evidence of deterioration in performance or unintended consequences, or any other concerns related to the measure. The Standing Committee may remove a measure from inactive endorsement status if the measure no longer meets NQF endorsement criteria. A maintenance review may occur upon a request from the Standing Committee or measure steward to return the measure to active endorsement.

Measures in reserve status will be considered for harmonization with related or competing measures. Measure developers should be aware of measures in reserve status and avoid developing duplicative measures.

Scientific Methods Panel: Frequently Asked Questions

Why did NQF create a Scientific Methods Panel?

In 2017, NQF underwent a redesign of its Consensus Development Process (CDP). This effort involved 50 stakeholders including representatives from NQF member organizations, the federal government, and NQF staff. One of the recommendations from that effort was to establish a Scientific Methods Panel (SMP) that would help ensure higher-level and more consistent evaluation of the scientific acceptability (i.e., reliability, validity) of complex measures, as well as encourage greater engagement and participation by consumers, patients, and purchasers on NQF standing committees.

What does the Scientific Methods Panel do?

The new panel has two specific charges:

- Evaluate complex measures for the criterion of scientific acceptability, with a focus on reliability and validity analyses and results.
- Serve in an advisory capacity to NQF on methodologic issues related to measure testing, risk adjustment, and emerging measurement approaches.

What expertise do you need to be a member of the Scientific Methods Panel?

The NQF SMP consists of up to 30 individuals with expertise in statistics, risk-adjustment, measure testing, psychometrics, economics, composite measures, and electronic clinical quality measures (eCQMs). It is co-led by NQF staff and two co-chairs designated by NQF. Each new panel member will serve an initial term of three years, with an optional two-year term to follow. The Consensus Standards Approval Committee (CSAC) oversees the work of the Scientific Methods Panel as part of its oversite of all of NQF's Consensus Development Process.

Is the Scientific Methods Panel a multistakeholder group?

Because the charge of the SMP is methodological in nature, NQF sought individuals with specific methodological expertise rather than those with particular stakeholder perspectives. While not quite as diverse as other NQF committees, the membership of the SMP does include academic and other researchers, healthcare providers, informaticists, consumers, and measure developers.

Does each NQF standing committee have its own Scientific Methods Panel?

There is only one SMP. It supports the standing committees for all 14 topical areas.

What defines a measure as complex or noncomplex?

The following types of measures are considered complex and therefore qualify for evaluation by the SMP:

- Outcome measures, including intermediate clinical outcomes
- Instrument-based measures (e.g., patient-reported, outcome-based performance measures)
- Cost/resource use measures
- Efficiency measures (those combining concepts of resource use and quality)
- Composite measures

Measures that do not fall under these categories are considered noncomplex (these typically are evaluated initially by NQF staff, then shared with standing committees). As part of their initial review of submitted measures, NQF staff identify and share with the SMP complex measures for evaluation.

How does the Scientific Methods Panel work?

Similar to the past work of NQF staff, the SMP provides NQF standing committees with evaluations and ratings of reliability and validity for new complex measures and for previously endorsed complex measures with updated testing. Standing committees consider this input when making their endorsement decisions. All panel members complete an annual, general disclosure of interest (DOI) form, as well as measure-specific disclosure forms to identify any need for recusal for specific measures.

Based on what was learned from previous evaluation cycles since Fall 2017, the SMP evaluation process has evolved. In the current process, panelists are assigned to evaluation subgroups; the number and size of the subgroups depends on the number of complex measures submitted for endorsement. Generally, each subgroup will comprise five to eight panel members. Each member conducts an in-depth evaluation of assigned measures. NQF staff assigns measures to subgroup members for evaluation based on panelists' relevant expertise, availability, and known disclosures. Subgroups discuss all measures deemed "consensus not reached" during a public meeting prior to voting. Subgroup members and staff also may request discussion and/or vote of other measures at will. Measure developers will be given the opportunity to provide additional information to the SMP prior to its final vote. The majority recommendations from the subgroup vote serve as the overall assessment of reliability and validity. The final results from the subgroup vote are shared with the appropriate standing committees, along with a summary of the SMP's evaluation. As per the current measure evaluation process, information about measures being evaluated will be posted on NQF's public webpages.

What is the process if the Scientific Methods Panel rates a measure as "low" or "insufficient" for reliability or validity?

Beginning the with the Fall 2019 evaluation cycle, measures rated by the SMP as "low" or "insufficient" for reliability or validity can be discussed by the relevant standing committee. The measure specifications, testing information, and a summary of the SMP's evaluation of these measures will be shared with the standing committee. Standing Committee members will have the option to pull measures that did not pass the SMP's evaluation for Committee discussion and, potentially, to revote on reliability and/or validity. Measures that do not pass the SMP evaluation and are not pulled by the relevant standing committee do not move forward in the process and will not be endorsed (if a new measure) or re-endorsed (if a maintenance measure). Measures that do not pass the SMP evaluation on reliability and/or validity, but are pulled by a Committee member for discussion, may be eligible for a revote. A measure is eligible for potential re-vote if it did not fail for one of the following reasons:

- Inappropriate methodology or testing approach applied to demonstrate reliability or validity
- Incorrect calculations or formulas used for testing
- Description of testing approach, results, or data is insufficient for SMP to apply the criteria
- Appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements

Measures that are not pulled for Committee discussion can be revised and resubmitted for reconsideration in a future cycle (there are 2 cycles per year). SMP evaluation summaries will be provided to the developer, and

therefore, any future resubmission can address the concerns of the Panel. NQF will inform the standing committee of the results of the SMP evaluation and the anticipated timing of resubmission.

Do the Scientific Methods Panel members provide the final vote for the Scientific Acceptability evaluation criterion?

No. The SMP will focus on issues related to methods and results of reliability and validity testing, as well as other methodological issues (e.g., statistical adequacy of risk- adjustment methodology). Their ratings will be provided as input for the standing committee's decision. It is possible that standing committees will have substantial clinical and other topical expertise to contribute to the evaluation of validity, in particular.

Will the standing committee vote on reliability and validity? What if it disagrees with the recommendations of Scientific Methods Panel?

If a standing committee agrees with the recommendations from the Panel regarding measures for which the Panel has rated as "moderate" or "high" for reliability and validity, and has no other concerns regarding the scientific acceptability of the measure (e.g., clinical perspectives that impact validity), it can accept the ratings provided by the SMP. Otherwise, the Committee will discuss their concerns and then vote on the criteria. Committee members can ultimately disagree with moderate or high recommendations and ratings provided by the SMP (or NQF staff). Measures that do not pass the SMP's evaluation can be pulled for further discussion and, potentially, for revote on reliability and/or validity, as described above.

Will the Scientific Methods Panel and standing committee review measures simultaneously?

Evaluation of complex measures by the SMP and the standing committee will not be simultaneous. The SMP will complete its evaluation of reliability and validity, and then NQF staff will complete the preliminary analysis for the remaining criteria. NQF staff will then collect all preliminary analyses for each topic area and forward those to developers for review. The developers will have at least 48 hours to review the preliminary analysis for factual accuracy. NQF staff will revise the preliminary analyses and recommendations, if needed, and then release all submission information, including the preliminary analyses and ratings from the SMP, to the appropriate standing committee for evaluation.

How will NQF ensure consistent evaluations by the Scientific Methods Panel?

NQF provides guidance documents for the SMP that are similar to those currently provided to standing committees. The guidance documents contain the SMP charge, terms and conditions, roles and responsibilities of panel members, and instructions on evaluating measures for scientific acceptability. Panel members will use the same algorithms for rating reliability and validity as used by standing committees. Panel members will use a templated worksheet to aid their evaluations. Further, NQF will convene the Panel bi-monthly to discuss methodological issues and how they should be considered relative to NQF's evaluation criteria.

What is the expected workload of Scientific Methods Panel members?

Using our knowledge of currently endorsed measures, past experience regarding the number, type, and complexity of new measures, and experience from prior Methods Panel evaluation cycles, NQF anticipates that each Panel member will evaluate the scientific acceptability of 15-30 measures per year (depending on availability, need for recusal, expertise, etc.). Panel members also will participate on bi-monthly webinars and two in-person meetings to evaluate measures, discuss methodologies and other testing-related issues, provide

guidance regarding these issues, and promote consistency in the evaluation of measures against NQF's endorsement criteria.

Will Scientific Methods Panel members be available during evaluation meetings to answer questions from the standing committee?

NQF will provide to the relevant standing committees the recommendations and rationale of the SMP on evaluated measures. Typically, panel members will not be available during standing committee evaluation meetings. Instead, NQF staff act as liaisons between the SMP and the standing committee. However, some Panel members are also standing committee members. In the event that the standing committee has a SMP member who evaluated a specific measure that is being evaluated by the standing committee, this person can discuss the measure and answer questions from the standing committee. However, the individual, as a member of the standing committee, will not be allowed to vote on the criteria of reliability and validity for that measure. The individual can vote on the other measure criteria.

If the Scientific Methods Panel only evaluates complex measures, how will noncomplex measures be evaluated?

Following the current process, NQF staff will evaluate noncomplex measures and provide preliminary ratings for reliability and validity. Standing Committees should consider these ratings as input to inform their endorsement decisions. The same process applies vis-à-vis forwarding measures to standing committees, as described above.