NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Guidance Summary Tables

Effective July 2013 (Updated 10/11/13)

Contents

Introduction

This document contains the measure evaluation criteria as well as additional guidance for evaluating the criteria. Additional information is available in detailed reports that can be accessed through NQF's Measure Evaluation web page.

Conditions for Consideration

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

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

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

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

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

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

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

Note

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

2. A measure that has not been tested for reliability and validity is only potentially eligible for time-limited endorsement if all of the following conditions are met: 1) the measure topic is not addressed by an endorsed measure; 2) it is relevant to a critical timeline (e.g., legislative mandate) for implementing endorsed measures; 3) the measure is not complex (requiring risk adjustment or a composite); and 4) the measure steward verifies that testing will be completed within 12 months of endorsement.

Criteria for Evaluation

If all conditions for consideration are met, measures are evaluated for their suitability based on standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Feasibility, Usability and Use*, and *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* or *Scientific Acceptability of Measure Properties*, 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, PRO-PMs, composite performance measures, eMeasures), 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); 2b3 (also composite); 2b4; 2b6; 4c (also composite); 5a and 5b (also composite),

1.Evidence, Performance Gap, and Priority (Impact)—Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.*** **Yes☐ No☐**

1a. Evidence to Support the Measure Focus H☐ M☐ L☐ I ☐

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

AND

1b. Performance Gap (see following evidence)

AND

1c. High Priority (see following evidence)

1d. For composite performance measures, the following must be explicitly articulated and logical: (see following evidence)

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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).

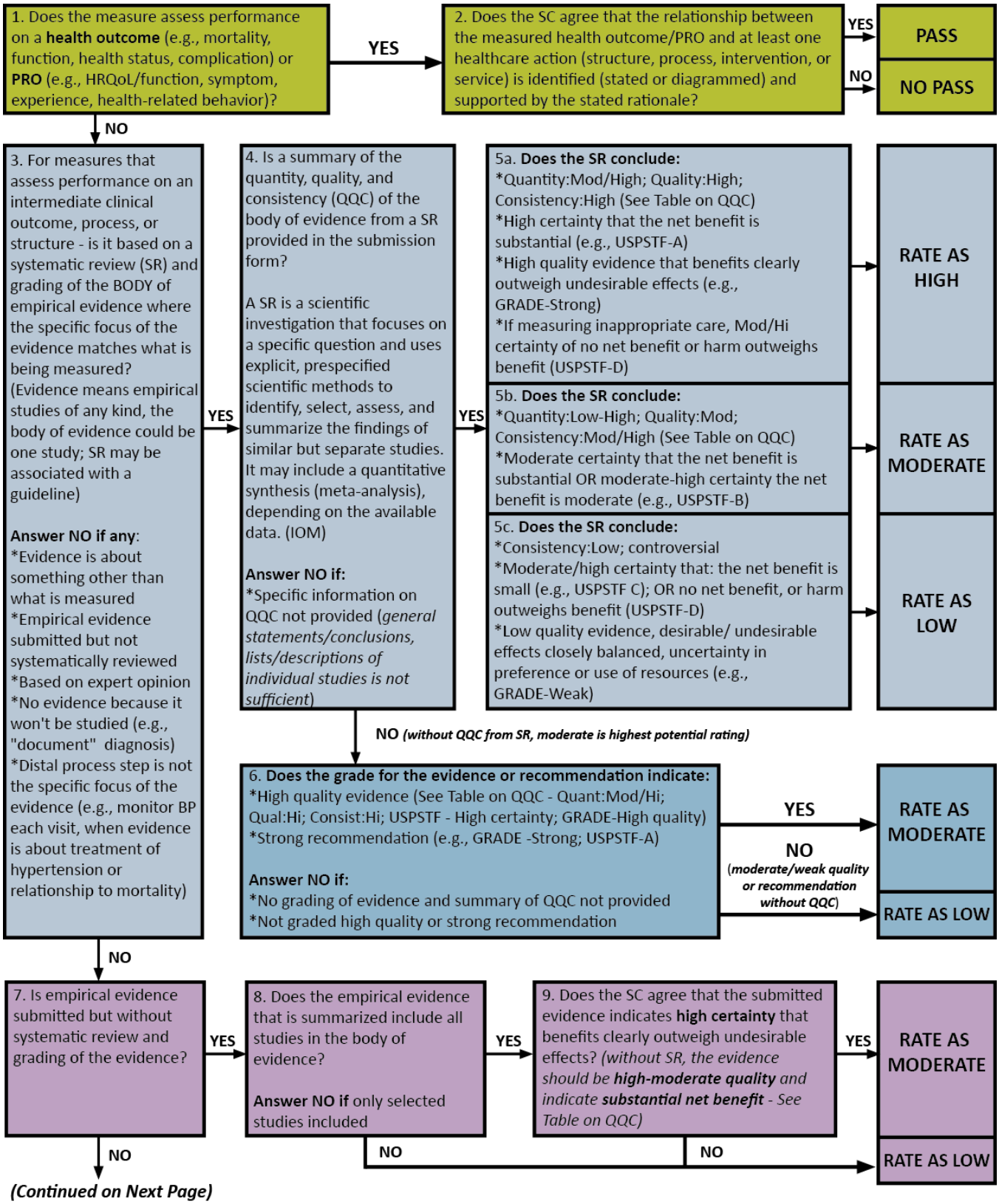
Guidance on Evaluating Importance to Measure and Report

Table 1: Evidence to Support the Focus of Measurement

TYPE OF MEASURE	EVIDENCE	EXAMPLE OF MEASURE TYPE AND EVIDENCE TO BE ADDRESSED
Health Outcome An outcome of care is the health status of a patient (or change in health status) resulting from healthcare— desirable or adverse. In some situations, resource use may be considered a proxy for a health state (e.g., hospitalization may represent deterioration in health status). **Patient-reported outcomes** include health-related quality of life/functional status, symptom/ symptom burden, experience with care, health-related behavior	A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service. See Table 5.	**#0230** Acute myocardial Infarction 30-day mortality Survival is a goal of seeking and providing treatment for AMI. **Rationale** healthcare processes/ interventions (aspirin, reperfusion) lead 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. **Rationale** healthcare processes (medication reconciliation, care coordination) lead to decreased hospitalization of patients receiving home care services **#0140** Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients Avoiding harm from treatment is a goal when seeking and providing healthcare. **Rationale** healthcare processes (ventilator bundle) lead to decreased ventilator acquired pneumonia **#0711** Depression remission at 6 months Relief of symptoms is a goal of seeking and providing healthcare services. **Rationale:** healthcare processes (use of antidepressants, psychotherapy) lead to decreased symptoms of depression #0166 HCAHPS experience with communication with doctors (assuming demonstration this is of value to patients) **Rationale:** healthcare practices (response time, respect, attention, explanation) leads to better experience with physician communication

TYPE OF MEASURE	EVIDENCE	EXAMPLE OF MEASURE TYPE AND EVIDENCE TO BE ADDRESSED
Intermediate Clinical Outcome An intermediate outcome is a change in physiologic state that leads to a longer-term health outcome.	Quantity, quality, and consistency of a body of evidence that the measured intermediate clinical outcome leads to a desired health outcome. See Table 4.	**#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 A process of care is a healthcare-related activity performed for, on behalf of, or by a patient.	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. See Table 4.	**#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 Structure of care is a feature of a healthcare organization or clinician related to its capacity to provide high-quality healthcare.	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). See Table 4.	**#0190** Nurse staffing hours **Evidence** that higher nursing hours result in lower mortality or morbidity, or leads to provision of effective care processes (e.g., lower medication errors) that lead to better outcomes

Algorithm 1. Guidance for Evaluating the Clinical Evidence



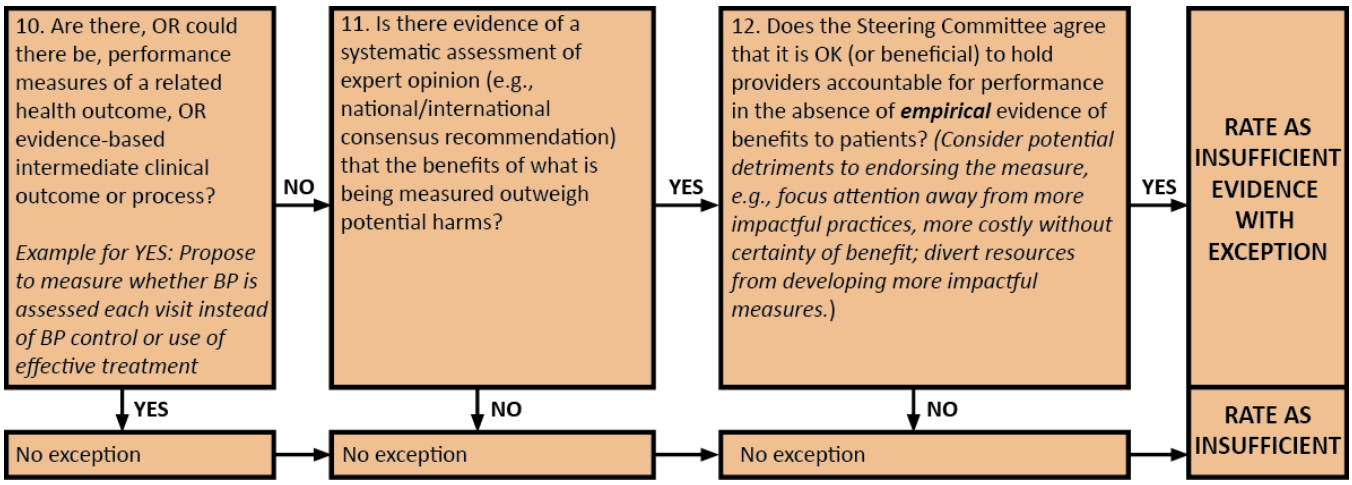

10. Are there, OR could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcome or process?
Example for YES: Propose to measure whether BP is assessed each visit instead of BP control or use of effective treatment
NO
11. Is there evidence of a systematic assessment of expert opinion (e.g., national/international consensus recommendation) that the benefits of what is being measured outweigh potential harms?
YES
12. Does the Steering Committee agree that it is OK (or beneficial) to hold providers accountable for performance in the absence of empirical evidence of benefits to patients? (Consider potential detriments to endorsing the measure, e.g., focus attention away from more impactful practices, more costly without certainty of benefit; divert resources from developing more impactful measures.)
YES
RATE AS INSUFFICIENT EVIDENCE WITH EXCEPTION
YES
No exception
NO
No exception
NO
No exception
RATE AS 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[a] 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[b]	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[b]	• 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)
Low	1 study[b]	• 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 *(See Table 3 for exceptions.)*	• 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

[a] *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.[15]
[b] The suggested number of studies for rating levels of quantity is considered a general guideline.

1.Evidence, Performance Gap, and Priority (Impact)—Importance to Measure and Report (continued)
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet <u>all subcriteria</u> to pass this criterion and be evaluated against the remaining criteria.*** **Yes☐ No☐**

1a. Evidence to Support the Measure Focus (see above)

AND

1b. Performance Gap H☐ M☐ L☐ I ☐

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

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

AND

1c. High Priority (previously referred to as High Impact) **H☐ M☐ L☐ I ☐**
The measure addresses:

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

OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).
- For patient-reported outcomes, there is evidence that the target population values the PRO and finds it meaningful.

1d. For composite performance measures, the following must be explicitly articulated and logical: **H☐ M☐ L☐ I ☐**

1d1. 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
1d2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
1d3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

Notes
7. Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

Table 3: Generic Scale for Rating Subcriteria 1b, 1c, 1d

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)

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.***
Yes☐ No☐

2a. Reliability H☐ M☐ L☐ I ☐
2a1. The measure is well defined and precisely specified [8] so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and must use the Quality Data Model (QDM) and value sets vetted through the National Library of Medicine's Value Set Authority Center (VASC). [9]

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b. Validity H☐ M☐ L☐ I ☐
2b1. The measure specifications [8] are consistent with the evidence presented to support the focus of measurement under criterion 1a. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score
.

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

AND

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

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance;
OR
there is evidence of overall less-than-optimal performance.

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

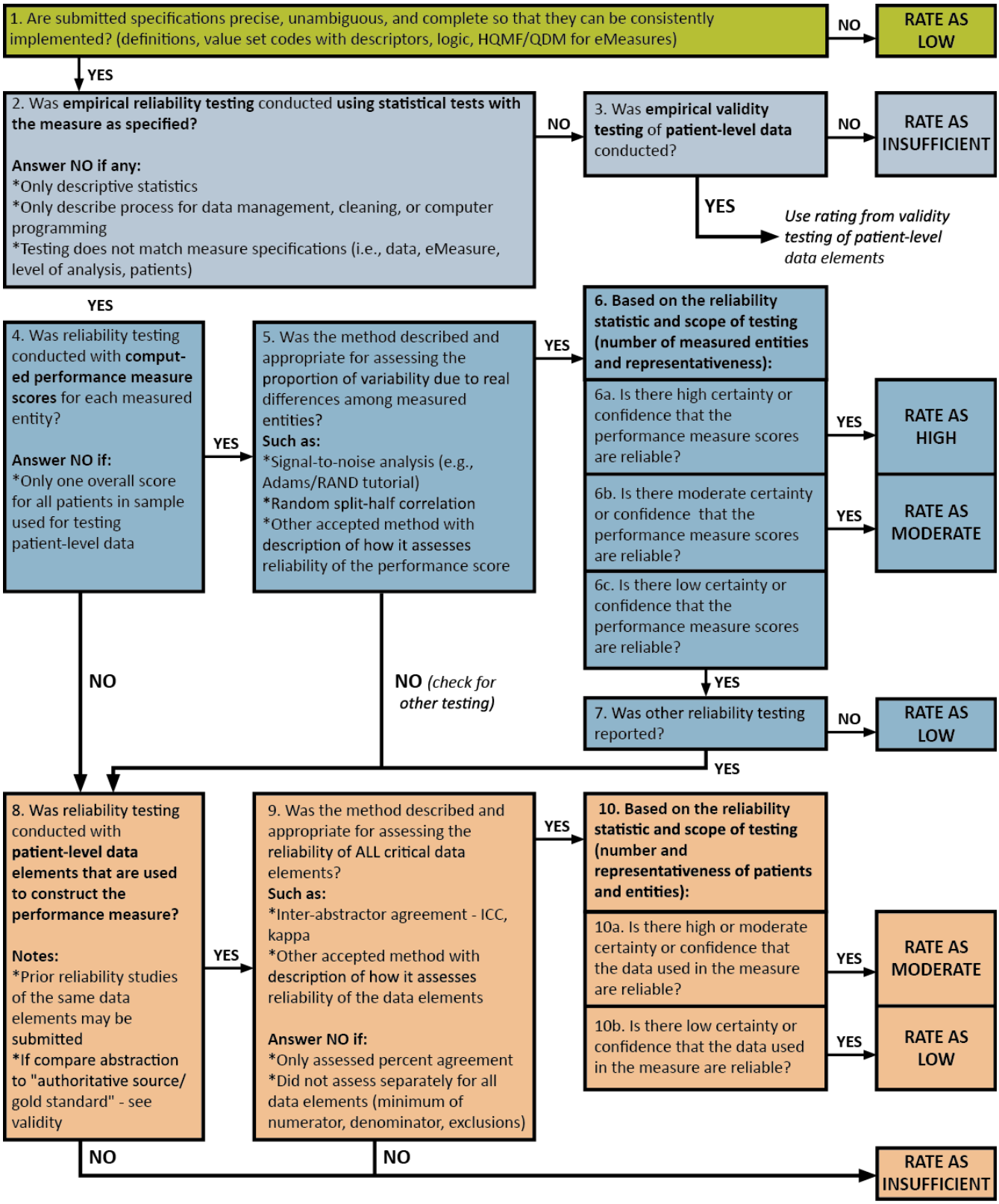
2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2c. Disparities *(Disparities should be addressed under subcriterion 1b)*
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);

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate the following: H☐ M☐ L☐ I ☐
2d1. 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
2d2. 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)*

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

Algorithm 2. Guidance for Evaluating Reliability (including eMeasures)



Algorithm 3. Guidance for Evaluating Validity (including eMeasures)

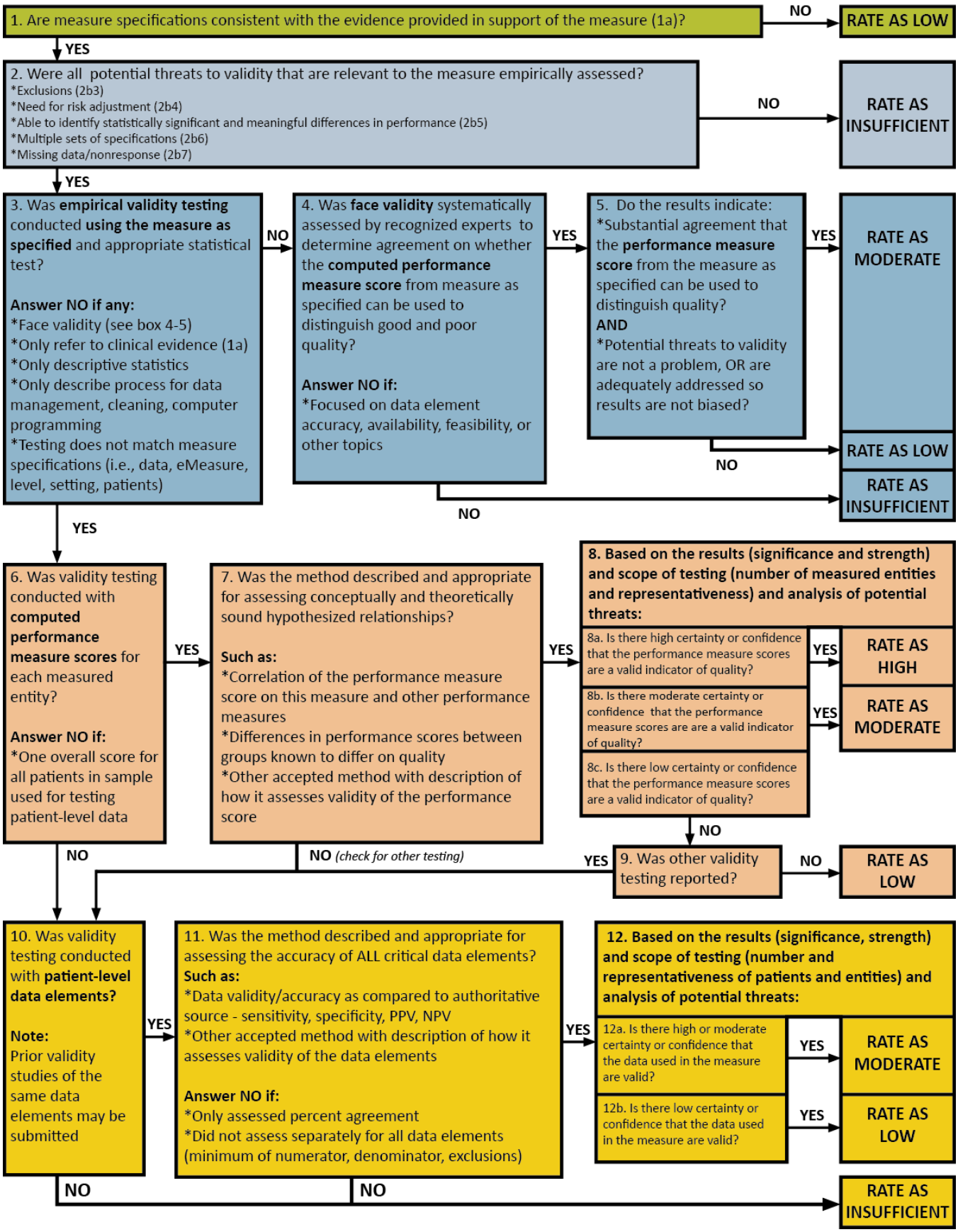


Table 4: Scope of Testing Required at the Time of Review for Endorsement Maintenance

	FIRST ENDORSEMENT MAINTENANCE REVIEW	SUBSEQUENT REVIEWS
Reliability	**Measure In Use** • Analysis of data from entities whose performance is measured • Reliability of measure scores (e.g., signal to noise analysis) **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 a high rating
Validity	**Measure in Use** • Analysis of data from entities whose performance is measured • Validity of measure score for making accurate conclusions about quality • 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 a high rating

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.
H☐ M☐ L☐ I ☐

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). **H☐ M☐ L☐ I ☐**

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. **H☐ M☐ L☐ I ☐**

3c. Demonstration that the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient-reported data, patient confidentiality, [17] 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). For eMeasures, a feasibility assessment addresses the data elements and measure logic [18] and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.
H☐ M☐ L☐ I ☐

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.
18. The feasibility assessment uses a standard score card or a fully transparent alternative that includes at a minimum: a description of the assessment, feasibility scores for all data elements, and explanatory notes for all data element components scoring a "1" (lowest rating); measure logic can be executed; with rationale and plan for addressing feasibility concerns.

Guidance on Evaluating Feasibility

Table 5: Generic Scale for Rating Feasibility Subcriteria

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 6. Data Element Feasibility Scorecard

DATA ELEMENT:			
Measure Title:			
Data element definition:			
Who performed the assessment:			
Type of setting or practice, i.e., solo practice, large group, academic hospital, safety net hospital, integrated system:			
EHR system used:			
	Current (1-3)	Future* (1-3)	Comments
Data Availability – Is the data readily available in structured format? Scale: 3 – Data element exists in structured format in this EHR. [2] – Not defined as this time. Hold for possible future use. 1 – Data element is not available in structured format in this EHR.			
Data Accuracy – Is the information contained in the data element correct? Are the data source and recorder specified? Scale: 3 – The information is from the most authoritative source and/or is highly likely to be correct. (e.g., laboratory test results transmitted directed from the laboratory information system into the EHR). 2 – The information may not be from the most authoritative source and/or has a moderate likelihood of being correct. (e.g., self-report of a vaccination). 1 – The information may not be correct. (e.g., a check box that indicates medication reconciliation was performed).			
Data Standards – Is the data element coded using a nationally accepted terminology standard? Scale: 3 – The data element is coded in nationally accepted terminology standard. 2 – Terminology standards for this data element are currently available, but is not consistently coded to standard terminology in the EHR, or the EHR does not easily allow such coding. 1 – The EHR does not support coding to the existing standard.			
Workflow – To what degree is the data element captured during the course of care? How does it impact the typical workflow for that user? Scale: 3 – The data element is routinely collected as part of routine care and requires no additional data entry from clinician solely for the quality measure and no EHR user interface changes. Examples would be lab values, vital signs, referral orders, or problem list entry. 2 – Data element is not routinely collected as a part of routine care and additional time and effort over and above routine care is required, but perceived to have some benefit. 1 – Additional time and effort over and above routine care is required to collect this data element without immediate benefit to care			
DATA ELEMENT FEASIBILITY SCORE			

*For data elements that score low on current feasibility, indicate the anticipated feasibility score in 3-5 years based on a projection of the maturation of the EHR, or maturation of its 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. **H☐ M☐ L☐ I ☐**

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

AND

4b. Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. [22] If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
H☐ M☐ L☐ I ☐

AND

4c. The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists). **H☐ M☐ L☐ I ☐**

Notes
19. Transparency is the extent to which performance results about identifiable, accountable entities are *disclosed and available* outside of the organizations or practices whose performance is measured. Maximal transparency is achieved with **public reporting** defined as making comparative performance results about identifiable, accountable entities freely available (or at nominal cost) to the public at large (generally on a public website). *At a minimum, the data on performance results about identifiable, accountable entities are available to the public (e.g., unformatted database)*. The capability to verify the performance results adds substantially to transparency.
20. This guidance is not intended to be construed as favoring measures developed by organizations that are able to implement their own measures (such as government agencies or accrediting organizations) over equally strong measures developed by organizations that may not be able to do so (such as researchers, consultants, or academics). Accordingly, measure developers may request a longer timeframe with appropriate explanation and justification.
21. Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.
22. 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.

Guidance on Evaluating Usability and Use

Table 7: Generic Scale for Rating Usability and Use Subcriteria

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 8. Key Questions for Evaluating Usability and Use

SUBCRITERIA	KEY QUESTIONS	SUITABLE FOR ENDORSEMENT?
3a, 3b, 3c	• Are all three subcriteria met? (3a—accountability/transparency, 3b—improvement, and 3c—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**
3a. 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 by three years? • Are the performance results publicly reported by 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 4a and/or 4b are 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 use in accountability application/public reporting (4a) and/or lack of improvement (4b); • credibility of plan for implementation for accountability/public reporting (4a) and/or credibility of rationale for improvement (4b); • 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 (4a) OR demonstration of improvement (4b) require judgment and supporting rationale.
3b. 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?	
3c. 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 4a and 4b.

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

5a. The measure specifications are harmonized [23] with related measures;
OR
the differences in specifications are justified.

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

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

Guidance on Evaluating Related and Competing Measures

Table 9: 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

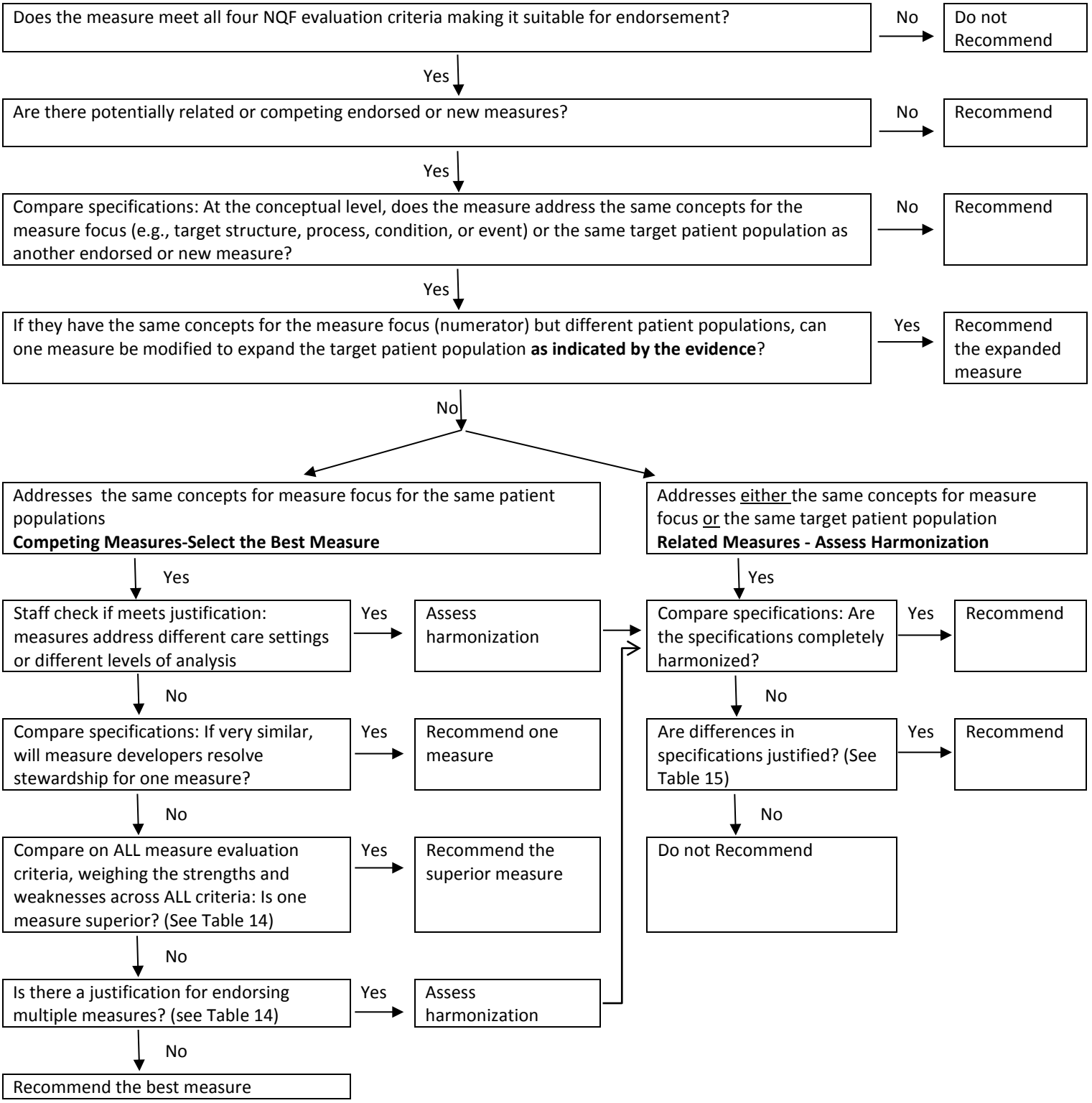


Table 10: Evaluating Competing Measures for Superiority or Justification for Multiple Measures

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 <u>weighing the strengths and weaknesses across ALL the criteria</u> 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, Priority—Importance to Measure and Report:** Competing measures generally will be the same in terms of the evidence for the focus of measurement (1a) and addressing a high-priority area of healthcare (1c) . However, due to differences in measure construction, they could differ on performance gap or alignment with national health goals/priorities or opportunity for improvement. • Compare measures on opportunity for improvement (1b) • Compare measures on alignment with specific national health goals/priorities (1c) **Reliability and Validity—Scientific Acceptability of Measure Properties:** • Compare evidence of reliability (2a1-2a2) • Compare evidence of validity, including threats to validity (2b1-2b7) 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 **<u>All else being equal on the criteria and subcriteria</u>**, <u>the preference is for</u>: • 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 **<u>All else being equal on the criteria and subcriteria</u>**, <u>the preference is for</u>: • 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. **<u>All else being equal on the criteria and subcriteria</u>**, <u>the preference is for</u>: • 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

STEPS	EVALUATE COMPETING MEASURES
	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.**
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 11: 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 population 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.

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 as measured by the PHQ-9 improved).	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 PRO-PMs

ABBREVIATED NQF ENDORSEMENT CRITERIA	CONSIDERATIONS FOR EVALUATING PRO-PMS THAT ARE RELEVANT TO OTHER PERFORMANCE MEASURES	UNIQUE CONSIDERATIONS FOR EVALUATING PRO-PMS
1. Importance to Measure and Report a. Evidence: Health outcome OR evidence-based intermediate outcome, process, or structure of care b. Performance gap c. High priority d. Composite	• PRO-PMs should have the same evidence requirement as health outcomes – rationale supports the relationship of the health outcome to processes or structures of care. • Exceptions to the evidence requirement for performance measures focused solely on administering a PROM should be addressed the same as other measures based solely on conducting an assessment (e.g., order lab test, check BP).	• Patients/persons must be involved in identifying PROs for performance measurement (person-centered; meaningful).
2. Scientific Acceptability of Measure Properties a. Reliability 1. Precise specifications 2. Reliability testing (data elements or performance measure score) b. Validity 1. Specifications consistent with evidence 2. Validity testing (data elements or performance measure score) 3. Exclusions 4. Risk adjustment 5. Identify differences in performance 6. Comparability of multiple sets of specifications 7. Missing data/non-response	• Data collection instruments (tools) should be identified (e.g., specific PROM instrument, scale, or single item). • If multiple data sources (i.e., PROMs, 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 (PROM) and the PRO-PM performance measure score. • Differences in individuals' PROM values related to PROM 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 PROMs, are minimized and do not outweigh the benefit of performance measurement.	• The burden to respondents (people providing the PROM data) should be minimized (e.g., availability and accessibility enhanced by multiple languages, methods, modes). • Infrastructure to collect PROM data and integrate into workflow and EHRs, as appropriate.
4. Usability and Use a. Accountability and transparency b. Improvement c. Benefits outweigh unintended negative consequences	• Adequate demonstration of the criteria specified above supports usability and ultimately the use of a PRO-PM for accountability and performance improvement.	

ABBREVIATED NQF ENDORSEMENT CRITERIA	CONSIDERATIONS FOR EVALUATING PRO-PMS THAT ARE RELEVANT TO OTHER PERFORMANCE MEASURES	UNIQUE CONSIDERATIONS FOR EVALUATING PRO-PMS
5. Comparison to Related or Competing Measures **5a.** Harmonization of related measures **5b.** Competing measures	• Apply to PRO-PMs	• PRO-PMs specified to use different PROM 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.

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

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

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

The following **will not be** considered composite performance 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 or greater than 3).
- Measures with multiple measure components that are assessed for each patient, but that result in multiple scores for an accountable entity, rather than a single score. These generally should be submitted as separate measures and indicated as paired/grouped measures.
- Measures of multiple linked steps in one care process assessed for each patient. These measures focus on one care process (e.g., influenza immunization) but may include multiple steps (e.g., assess immunization status, counsel patient, and administer vaccination). These are distinguished from all-or-none composites that capture multiple care processes or outcomes (e.g., foot care, eye care, glucose control).
- Performance measures of one concept (e.g., mortality) specified with a statistical method or adjustment **(e.g., empirical Bayes shrinkage estimation)** that combines information from the accountable entity with information on average performance of all entities or a specified group of entities (e.g., by case volume), **typically in order to increase reliability**.

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

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. High priority d. 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). The priority criterion (1c) applies to the composite performance measure as a whole. 1d. 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. Specifications consistent with evidence 2. Validity testing (data elements or performance measure score) 3. Exclusions 4. Risk adjustment 5. Identify differences in performance 6. Comparability of multiple sets of specifications 7. Missing data/non-response 2c. Disparities 2d. 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	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. **2b2.** For composite performance measures, validity should be empirically demonstrated for the composite measure score. If empirical testing is not feasible at the time of initial endorsement, acceptable alternatives include systematic assessment of content or face validity of the composite performance measure or demonstration that each of the component measures meet NQF subcriteria for validity. By the time of endorsement maintenance, validity of the composite performance measure must be empirically demonstrated.

ABBREVIATED NQF ENDORSEMENT CRITERIA	GUIDANCE FOR COMPOSITE PERFORMANCE MEASURES
specified handling of missing data minimizes bias (i.e., achieves scores that are an accurate reflection of quality).	It is unlikely that a "gold standard" criterion exists, so validity testing generally will focus on construct validation – testing hypotheses based on the theory of the construct. Examples include testing the correlation with measures hypothesized to be related or not related; testing the difference in scores between groups known to differ on quality assessed by some other measure. **2b3.** Applies to the component measures and composite performance measures. **2b4.** Applies to outcome component measures (unless NQF-endorsed). **2b5.** Applies to composite performance measures. **2b6.** Applies to component measures. **2b7.** 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.** Applies to composite performance measures. 2d. 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 [15], 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.

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 a. Accountability and transparency b. Improvement c. 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). **4a.** Applies to composite performance measures. To facilitate transparency, at a minimum, the individual component measures of the composite must be listed with use of the composite measure. **4b.** Applies to composite performance measures. **4c.** Applies to composite performance measures and component measures. If there is evidence of unintended negative consequences for any of the components, the developer should explain how that is handled or justify why that component should remain in the composite.
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.