NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Guidance Summary Tables
Effective for Projects Beginning after January 2011

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**□ Guidance-Table 7

2a. Reliability **H**□ **M**□ **L**□ **I** □ Guidance-Table 6; EHR measures-Table 8
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]

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.

2b. Validity **H**□ **M**□ **L**□ **I** □ Guidance-Table 6; EHR measures-Table 8
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.

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.

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

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]

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.

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.

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

2c. Disparities H□ M□ L□ I □ Definitions-Table 9
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.

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.
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).
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 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.
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.
12. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
13. Risk factors that influence outcomes 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.
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 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.

NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Guidance Summary Tables
Effective for Projects Beginning after January 2011

Guidance on Evaluating Scientific Acceptability of Measure Properties

For more information, see full report: Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

Table 6: Evaluation Ratings for Reliability and Validity

Rating	Reliability	Validity
High	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Empirical evidence of reliability of **BOTH** data elements (Table A-2) **AND** measure score (Table A-1) within acceptable norms: • Data element: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity (Table A-4) was demonstrated*; **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of **BOTH** data elements (Table A-4) **AND** measure score (Table A-3) within acceptable norms: • Data element: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or “incorrect” data) are empirically assessed and adequately addressed so that results are not biased
Moderate	All measure specifications are unambiguous as noted above **AND** Empirical evidence of reliability within acceptable norms for either critical data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either critical data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased

Low	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in identifying who is included and excluded from the target population, or the event, condition, or outcome being measured; or how to compute the score, etc.; **OR** Empirical evidence (using appropriate method and scope) of unreliability for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results
Insufficient Evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above); **OR** Threats to validity as noted above are likely and are NOT empirically assessed

Table 7: Evaluation of Scientific Acceptability of Measure Properties Based on Reliability and Validity Ratings

Validity Rating	Reliability Rating	Pass *Scientific Acceptability of Measure Properties* for Initial Endorsement*	
High	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

*A measure that does not pass the criterion of *Scientific Acceptability of Measure Properties* would not be recommended for endorsement.

Measure Evaluation Criteria and Guidance Summary Tables
Effective for Projects Beginning after January 2011

Table 8: Evaluation of Reliability and Validity of Measures Specified for EHRs

	New Measure Specified for EHR		Modifications for Endorsed Measure *Re-specified* for EHRs
Rating	Reliability Description and Evidence	Validity Description and Evidence	
High	All EHR measure specifications are unambiguous[+] and include only data elements from the Quality Data Model (QDM)***** including quality data elements, code lists, and measure logic; **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability of both data element **AND** measure score within acceptable norms: • Data element: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) reflect the quality of care problem (1a,1b) and evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of both data elements **AND** measure score within acceptable norms: • Data element: validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the entire EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the Quality Data Model (QDM)***** and include quality data elements, code lists, and measure logic; **AND** Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality; **AND** Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits
Moder-ate	All EHR measure specifications are unambiguous[+] and include only data elements from the QDM;* **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability within acceptable norms for either data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the QDM as noted above **AND** Crosswalk of the EHR measure specifications as noted above demonstrates that they represent the original measure **AND** For measures with time-limited status, testing of the original measure and evidence ratings of moderate for reliability and validity as described in Table 2.
Low	One or more EHR measure specifications are ambiguous[+] or do not use data elements from the QDM*; **OR** Empirical evidence of unreliability for either data elements **OR** measure score—i.e., statistical results	The EHR measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** measure score— i.e., statistical results outside of acceptable norms **OR**	The EHR measure specifications do not use only data elements from the QDM; **OR** Crosswalk of the EHR measure specifications as noted above identifies that they do not represent the original measure **OR**

	New Measure Specified for EHR		**Modifications for Endorsed Measure *Re-specified* for EHRs**
Rating	**Reliability Description and Evidence**	**Validity Description and Evidence**	
	outside of acceptable norms	Identified threats to validity noted above are empirically assessed and determined to bias results	For measures with time-limited status, empirical evidence of low reliability or validity for original time-limited measure
Insufficient evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above) **OR** Threats to validity as noted above are likely and are NOT empirically assessed	Crosswalk of the EHR measure specifications as noted above was not completed OR For measures with time-limited status, inappropriate method or scope of reliability or validity testing for original time-limited measure

[+] Specifications are considered unambiguous if they are likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.

*QDM (formerly called the QDS) elements should be used when available. When quality data elements are needed but are not yet available in the QDM, they will be considered for addition to the QDM.

Table 9: Generic Scale for Rating Subcriterion 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)