NATIONAL
QUALITY FORUM

Measure Evaluation Criteria Refresher

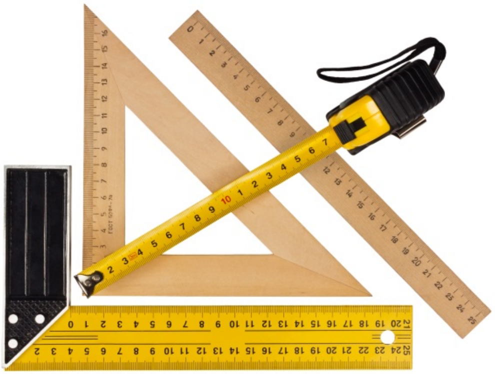
CSAC Informational Update

April 24, 2019

Overview of Presentation

- Provide an overview of NQF's measure evaluation criteria

Why Do We Measure?



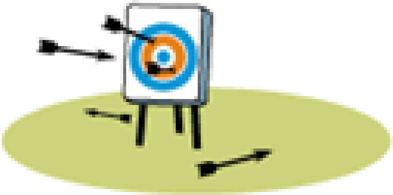
The primary goal of healthcare performance measurement is to **improve the quality (and access and cost) of healthcare** received by patients

(and ultimately, to **improve health**)

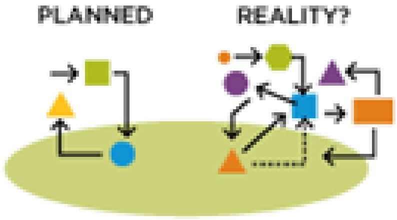
Measurement is a quality improvement tool, not an end in and of itself

What Makes a Great Measure?
NQF's Measure Evaluation Criteria

1 IMPORTANCE TO MEASURE AND REPORT

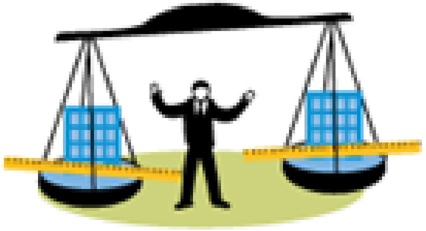


2 SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES



3 FEASIBILITY

4 USABILITY AND USE



5 ASSESS RELATED AND COMPETING MEASURES

NQF Measure Evaluation Criteria for Endorsement

NQF endorses measures that are suitable for both quality improvement efforts as well as for accountability applications (public reporting, payment programs, accreditation, etc.)

- Standardized evaluation criteria
- Criteria have evolved over time in response to stakeholder feedback
- The quality measurement enterprise is constantly growing and evolving—greater experience, lessons learned, expanding demands for measures—the criteria evolve to reflect the ongoing needs of stakeholders

Major Endorsement Criteria

- **Importance to measure and report**: Goal is to measure those aspects with greatest potential of driving improvements; if not important, the other criteria are less meaningful **(must-pass)**
- **Reliability and Validity-scientific acceptability of measure properties**: Goal is to make valid conclusions about quality; if not reliable and valid, there is risk of improper interpretation **(must-pass)**
- **Feasibility**: Goal is to, ideally, cause as little burden as possible; if not feasible, consider alternative approaches
- **Usability and Use** (Use is **must-pass** for maintenance measures): Goal is to use for decisions related to accountability and improvement, to achieve high-quality, efficient healthcare for individuals or populations
- **Comparison to related or competing measures**

Criterion #1: Importance to Measure and Report

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

1a. Evidence: the measure focus is evidence-based

1b. Opportunity for Improvement: 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. Quality construct and rationale (composite measures only)

Subcriteron 1a: Evidence

- Outcome measures
  - *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*
- Structure, process, intermediate outcome measures
  - *The quantity, quality, and consistency of the body of evidence underlying the measure should demonstrate that the measure focuses on those aspects of care known to influence desired patient outcomes*
    - Empirical studies (expert opinion is not evidence)
    - Systematic review and grading of evidence preferred
- For measures derived from “patient” report
  - *Evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful*
  - *Current requirements for structure and process measures also apply to patient-reported structure/process measures*

Key Points for Evaluating Evidence

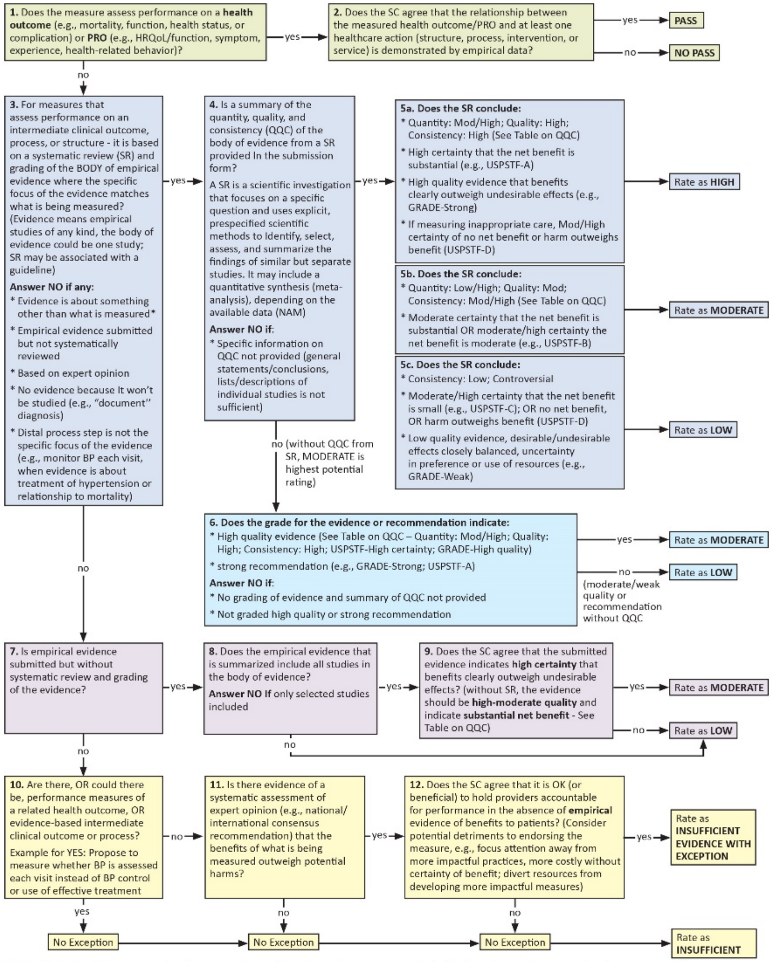
- The evaluation of the evidence subcriterion depends on the type of measure under consideration.
- Evidence should be presented about the relevant body of evidence—not selected individual studies.
- Ideally, measure developers will summarize a systematic review of the evidence that has been assembled, reviewed, and graded by others.
- Expert opinion is not considered empirical evidence, but evidence is not limited to randomized controlled trials.
- Measures with inconsistent or conflicting evidence should not pass the evidence subcriterion.

Key Points for Evaluating Evidence

- When evaluating the quality of the evidence, consider the following:
  - *The study design itself (e.g., RCT, non-RCT) or flaws in the design or conduct of the study (e.g., lack of blinding; large losses to follow-up)*
  - *The directness/indirectness of the evidence to the measure as specified (e.g., regarding the population, intervention, comparators, and/or outcomes)*
  - *Imprecision in study results (i.e., wide confidence intervals due to few patients or events)*
- Under limited circumstances, an exception to the evidence subcriterion may be invoked and evaluated according to the evidence algorithm

Algorithm for Rating Evidence

Algorithm 1. Guidance for Evaluating the Clinical Evidence



1. Does the measure assess performance on a **health outcome** (e.g., mortality, function, health status, or complication) or **PRO** (e.g., HRQoL/function, symptom, experience, health-related behavior)?

yes → **2.** Does the SC agree that the relationship between the measured health outcome/PRO and at least one healthcare action (structure, process, intervention, or service) is demonstrated by empirical data?

yes → **PASS**

no → **NO PASS**

no → **3.** For measures that assess performance on an intermediate clinical outcome, process, or structure - it is based on a systematic review (SR) and grading of the BODY of empirical evidence where the specific focus of the evidence matches what is being measured? (Evidence means empirical studies of any kind, the body of evidence could be one study; SR may be associated with a guideline)

Answer NO if any:

* Evidence is about something other than what is measured*

* Empirical evidence submitted but not systematically reviewed

* Based on expert opinion

* No evidence because It won't be studied (e.g., "document" diagnosis)

* Distal process step is not the specific focus of the evidence (e.g., monitor BP each visit, when evidence is about treatment of hypertension or relationship to mortality)

yes → **4.** Is a summary of the quantity, quality, and consistency (QQC) of the body of evidence from a SR provided In the submission form?

A SR is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to Identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data (NAM)

Answer NO if:

* Specific information on QQC not provided (general statements/conclusions, lists/descriptions of individual studies is not sufficient)

yes → **5a. Does the SR conclude:**

* Quantity: Mod/High; Quality: High; Consistency: High (See Table on QQC)

* High certainty that the net benefit is substantial (e.g., USPSTF-A)

* High quality evidence that benefits clearly outweigh undesirable effects (e.g., GRADE-Strong)

* If measuring inappropriate care, Mod/High certainty of no net benefit or harm outweighs benefit (USPSTF-D)

→ Rate as **HIGH**

5b. Does the SR conclude:

* Quantity: Low/High; Quality: Mod; Consistency: Mod/High (See Table on QQC)

* Moderate certainty that the net benefit is substantial OR moderate/high certainty the net benefit is moderate (e.g., USPSTF-B)

→ Rate as **MODERATE**

5c. Does the SR conclude:

* Consistency: Low; Controversial

* Moderate/High certainty that the net benefit is small (e.g., USPSTF-C); OR no net benefit, OR harm outweighs benefit (USPSTF-D)

* Low quality evidence, desirable/undesirable effects closely balanced, uncertainty in preference or use of resources (e.g., GRADE-Weak)

→ Rate as **LOW**

no (without QQC from SR, MODERATE is highest potential rating) → **6. Does the grade for the evidence or recommendation indicate:**

* High quality evidence (See Table on QQC – Quantity: Mod/High; Quality: High; Consistency: High; USPSTF-High certainty; GRADE-High quality)

* strong recommendation (e.g., GRADE-Strong; USPSTF-A)

Answer NO if:

* No grading of evidence and summary of QQC not provided

* Not graded high quality or strong recommendation

yes → Rate as **MODERATE**

no (moderate/weak quality or recommendation without QQC → Rate as **LOW**

no → **7.** Is empirical evidence submitted but without systematic review and grading of the evidence?

yes → **8.** Does the empirical evidence that is summarized include all studies in the body of evidence?

Answer NO If only selected studies included

yes → **9.** Does the SC agree that the submitted evidence indicates **high certainty** that benefits clearly outweigh undesirable effects? (without SR, the evidence should be **high-moderate quality** and indicate **substantial net benefit** - See Table on QQC)

yes → Rate as **MODERATE**

no → Rate as **LOW**

no → Rate as **LOW**

no → **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

yes → No Exception

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?

no → No Exception

yes → **12.** Does the SC 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**

no → No Exception

No Exception → No Exception → No Exception → Rate as **INSUFFICIENT**

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.

Subcriteron 1b: Opportunity for Improvement

- Underlying question: is there a quality problem?
- We expect current data for the measure as specified; (relevant data from the literature also may be used, especially for initial endorsement)
- When evaluating whether there is opportunity for improvement, consider:
  - *The distribution of performance scores*
  - *The number and representativeness of the entities included in the measure performance data*
  - *The size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem*
  - *Data on disparities*
- Topped out measures: potential for Reserve Status

Criterion #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 healthcare delivery

2a. Reliability (must-pass)

2a1. Precise specifications

2a2. Reliability testing—data elements or measure score

2b. Validity (must-pass)

2b1. Validity testing—data elements or measure score

2b2. Justification of exclusions—relates to evidence

2b3. Risk adjustment—typically for outcome/cost/resource use

2b4. Identification of differences in performance

2b5. Comparability of data sources/methods

2b6. Missing data

2c. Analysis support composite construction approach (must-pass)

Questions These Subcriteria Address

Reliability

- Are the specifications clear so that everyone will calculate the measure in the same way?
- Is the variation between providers primarily due to real differences in performance? Or is it because there is a lot of "noise" in the measurement?

Validity

- Is the measure actually measuring what it is intended to measure (e.g., quality of care)?
- Do the results of the measurement allow for correct conclusions about quality of care?

Key Points for Evaluating Scientific Acceptability

- Scientifically acceptable measures must be both reliable and valid
- Empirical demonstration of reliability and validity is expected, although for new measures, demonstration of face validity of the measure score as an indicator of quality also is allowed
- NQF is not prescriptive about how empirical measure testing is done; similarly, NQF does not set minimum thresholds for reliability or validity testing results
- Reliability and validity must be demonstrated for the measure as specified (including data source and level of analysis)

Key Points for Evaluating Scientific Acceptability

- Depending on the measure type, NQF may allow testing at either the data element level (using patient-level data) or at the performance measure score level (using data that have been aggregated across providers).
- When evaluating measure testing results, the method of testing, the data used for testing (often from a sample), and the results of the testing must be considered
- All three subcriteria under Scientific Acceptability are "must-pass"; therefore, each must be met in order to be recommended for endorsement

Some Additional Considerations

- We have rating algorithms for reliability and validity
- Most criteria/subcriteria involve a matter of degree rather than all-or-nothing determination—this requires both evidence and expert judgment
- Reliability
  - *The foundation for reliability is good specifications: definitions, codes, and instructions on how to calculate the measure*
  - *Evaluating testing for reliability can be tricky—several methods available, but no thresholds for results*
  - *Even so, assessing reliability is likely less subjective than assessing validity*
- Validity
  - *Assessing threats to validity even more important than testing*

Some Additional Considerations: Accounting for Social Risk Factors

The Standing Committee will be asked to consider the following questions:

- Is there a conceptual relationship between the SDS factor and the measure focus?
- What are the patient-level sociodemographic variables that were available and analyzed during measure development?
- Does empirical analysis (as provided by the measure developer) show that the SDS factor has a significant and unique effect on the outcome in question?
- Does the reliability and validity testing match the final measure specifications?

Criterion #3: Feasibility

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

3a: Clinical data generated during care process
3b: Electronic sources
3c: Data collection strategy can be implemented

- This is not a must-pass criterion
  - *HOWEVER, feasibility is critical for eCQMs*

Criterion #4: Usability and Use

Extent to which potential audiences 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.

Use (4a) Must-pass for maintenance measures

***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*

***4a2: Feedback by those being measured or others**: Those being measured have been given results and assistance in interpreting results; those being measured and others have been given opportunity for feedback; the feedback has been considered by developers*

Usability (4b)

***4b1: Improvement**: Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.*

***4b2: Benefits outweigh the harms**: 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)*

Criterion #5: Related or Competing Measures

If a measure meets the four criteria and there are endorsed/new **related** measures (same measure focus or same target population) or **competing** measures (both the same measure focus and 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.
- 5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure) **OR** multiple measures are justified.