TO: Consensus Standards Approval Committee

FR: Reva Winkler and Karen Johnson

RE: Potential changes to NQF's measure evaluation criteria and/or guidance

DA: July 6, 2016

During measure evaluation discussions, Committees often ask questions or provide feedback on NQF's measure evaluation criteria and guidance for Committees. Recent Committee discussions have raised several issues that indicate a possible need for change to NQF's measure evaluation criteria and/or guidance.

**CSAC ACTION REQUIRED**

The CSAC will review, discuss, and/or approve (as needed) proposed changes to NQF's endorsement evaluation criteria and/or guidance.

**PROPOSED CHANGE TO EVIDENCE REQUIREMENT FOR OUTCOME MEASURES**

The evidence requirement for health outcome measures is a rationale for how the outcome is influenced by at least one healthcare process or structure. The current requirement for a rationale is minimal and some Committee members believe that outcome measures should not get a "pass" on the evidence criterion.

To address this concern, we request the CSAC's feedback on revising the evidence requirement for outcome measures to specify some empirical evidence rather than just a rationale.

Potential change for discussion:

Criterion 1a. ~~A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.~~

At least one empirical study demonstrates an evidenced-based relationship between the health outcome and a healthcare structure, process, intervention, or service.

and

in the algorithm, change Box 2 wording:

> 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 identified (stated or diagrammed) and supported by ~~the stated rationale~~ at least one empirical study?

**PROPOSED CHANGE TO THE GUIDANCE FOR COMMITTEES IN EVALUATING OPPORTUNITY FOR IMPROVEMENT/GAP FOR MORTALITY AND PATIENT SAFETY EVENTS**

Criterion 1b. Opportunity for improvement/gap in care addresses the question of whether there is a quality problem that can be addressed by measurement. Data should demonstrate that there is an opportunity for improvement (i.e., overall poor performance, substantial variation across providers,

or variation for subpopulations (disparities in care)). Committees struggle with applying this criterion for low incidence patient safety events (that are usually very low but represent significant quality problems) or mortality measures that will never reach 0% but it is unknown what "perfect performance" might be.

To clarify direction to the Committees, we propose the following language (in red below) be added to guidance materials:

**1b. Performance Gap**

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.

When assessing measure performance data for Performance Gap (1b), the following factors should be considered:

- distribution of performance scores;
- number and representativeness of the entities included in the measure performance data;
- data on disparities; and
- size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.

Proposed addition:

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

**PROPOSAL FOR CRITERIA TO SUPPORT THE FINAL VOTE ON OVERALL RECOMMENDATION FOR ENDORSEMENT**

During the evaluation of a measure, the Committee votes on each major criterion and/or subcriterion and then votes on "Overall Recommendation for Endorsement." Measure developers have questioned this final vote, arguing that if a measure passes all the criteria it should be endorsed automatically. There have been instances where a measure is not recommended as suitable for endorsement, even though the measure has passed all the criteria. Often staff cannot understand or explain the rationale for what may appear to be an inconsistent vote.

Question for CSAC: Should NQF develop criteria for the final vote of "Overall Recommendation for Endorsement?" Criteria might include alignment with NQF priorities (to be identified for the Strategic Plan); filling a NQF-identified gap; potential impact on health of patients (i.e., likelihood of moving the quality needle).

**PROPOSED CHANGE TO GUIDANCE FOR IDENTIFYING COMPOSITE MEASURES FOR THE PURPOSES OF NQF MEASURE SUBMISSION, EVALUATION, AND ENDORSEMENT**

In 2012, an NQF-convened Technical Expert Panel (TEP) defined a composite performance measure as *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.* The TEP provided explicit guidance to clearly delineate what types of measures will be considered by NQF to be composite performance measures, as follows:

- 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).

All-or-none measures assess whether all essential care processes were received, or all outcomes were experienced, by each patient. Any-or-none measures assess whether any of a list of adverse outcomes were experienced, or inappropriate or unnecessary care processes were received by each patient. Although not unanimous, a majority of the TEP agreed that all-or-none and any-or-none measures should be considered composite performance measures. These measures are similar in construction in that all the components are assessed separately for each patient.

In addition to the usual evaluation criteria, composite measures must have a logical quality construct and rationale and empirical analyses to support the composite construction approach. Also, several of the other measure evaluation criteria differ somewhat for composite measures (e.g., reliability must be demonstrated for the composite measure score, meaning that demonstration of the reliability of the data elements used in the components of the composite is not sufficient).

While achieving overall consensus regarding the evaluation criteria and guidance, the TEP was not unanimous in its recommendation to identify **any-or-none** measures as composite performance measures. The TEP discussed whether any-or-none measures that include a group of patient-specific outcomes, such as complications, should always be considered composites. For surgical patients, for example, a developer may create a measure that looks for various events that may occur as unintended consequences of the operation for each patient. These measures may include events that have not previously been considered as individual measures (e.g., hemorrhage) or events that have previously been considered as individual measures (e.g., death, readmission). In some instances, the developer may not view a measure that incorporates multiple events such as complications as an any-or-none composite (e.g., complications are viewed as a single measure instead of multiple measures).

Ultimately, the CSAC agreed that such measures will be considered composites, with the expectation that the information needed to evaluate the composite-specific criteria is provided. However, if the developer provides a conceptual justification as to why such a measure should not be considered a composite, and that justification is accepted by the NQF committee, the measure can then be considered a single measure rather than a composite.

Because the TEP did not achieve unanimity in its recommendations, it recommended a review of its decisions after gaining more experience with composites. In the three years since the TEP's recommendations, there have been relatively few composite measures submitted to NQF. In general, both measure developers and Committees have been able to respond to/apply the additional criteria for composite measures. However, there has been almost universal push-back from developers in identifying **any-or-none** measures as composite measures. Even though the TEP allowed for an "exception" for these types of measures if justified by the developer, ensuring that the developer has utilized the correct submission form has been problematic and inconsistent, and explaining the nuances to Committees has engendered confusion and inconsistencies across projects and measures. In contrast, while there has been some resistance in identifying all-or-none measures as composite performance measures, in general, developers and Committees have complied with this guidance.

Potential change for discussion: No longer require **any-or-none** measures to be identified as composite performance measures for the purposes of NQF measure submission, evaluation, and endorsement.

# Potential changes to NQF's measure evaluation criteria and/or guidance

CSAC Meeting – July 13-14, 2016

NATIONAL QUALITY FORUM

# Evidence requirement for outcome measures (Criterion 1a)

- Current: A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service.
- Potential change for discussion: At least one empirical study demonstrates an evidenced-based relationship between the health outcome and a healthcare structure, process, intervention, or service.
  - To be accompanied by corresponding change in evidence algorithm

# Guidance on Opportunity for Improvement for mortality and patient safety measures

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

# Criteria to support final vote on overall recommendation for endorsement

- Question for CSAC discussion: Should NQF develop criteria for the final vote of "Overall Recommendation for Endorsement?" Criteria might include:
  - alignment with NQF priorities (to be identified for the Strategic Plan)
  - filling a NQF-identified gap
  - potential impact on health of patients (i.e., likelihood of moving the quality needle)

# Guidance for identifying composite measures

- Potential change :  No longer require **any-or-none** measures to be identified as composite performance measures for the purposes of NQF measure submission, evaluation, and endorsement

# NQF Measure Evaluation Criteria

July 2016

NQF's measure evaluation criteria—which reflect desirable characteristics of performance measures—are used to determine the suitability of measures for endorsement. NQF endorsement is intended to identify those performance measures that are most likely to facilitate achievement of high quality and efficient healthcare for patients. The criteria were originally established by the Strategic Framework Board[1] at NQF's origins. Experience with use of the criteria and feedback from NQF members and other users have prompted revisions and updates over the years to both the criteria and to the guidance that NQF issues for applying the criteria. CSAC oversees the consensus development process (CDP) and the measure evaluation process, including the criteria. Updates or revisions to the criteria are made in response to issues that arise during the consensus development process.

The ordering of the criteria and subcriteria is deliberate, as is the designation of some criteria and subcriteria as "must-pass."

**Criterion 1. Importance to measure and report** (must-pass) reflects the goal of measuring those aspects with greatest potential of driving improvements. Specifically, measures that are "Important to Measure and Report" are evidence-based and reflect variation in performance, overall less-than-optimal performance, or disparities. This criterion allows for a distinction between things that are important to do in clinical practice versus those that rise to the level of importance required for a national performance measure.

> 1a. Evidence to Support the Measure Focus (must-pass) considers whether there is adequate empirical evidence to support a measure for use as a national consensus standard. Evidence should establish a relationship to patient outcomes. Expert opinion is not considered to be empirical evidence.
>
> > *Process, Intermediate Outcome, Structure measures*: The strength of evidence is determined by the **quantity**, **quality**, and **consistency** of studies from the relevant body of evidence that relates the measure focus to health outcomes.
> >
> > *Health outcome measures and patient-reported outcome performance measures*: Evidence requires a rationale (which often includes studies) for how the outcome is influenced by healthcare processes or structures.
> >
> > *Exceptions to the evidence criterion*: An exception may be granted if the Committee/NQF agree that it is acceptable (or beneficial) to hold providers accountable for performance in the absence of empirical evidence of benefits to patients. Exceptions to evidence should occur infrequently.

[1] McGlynn EA, Selecting common measures of quality and system performance. *Med Care* 2003 Jan;41 (1 Suppl):139-47

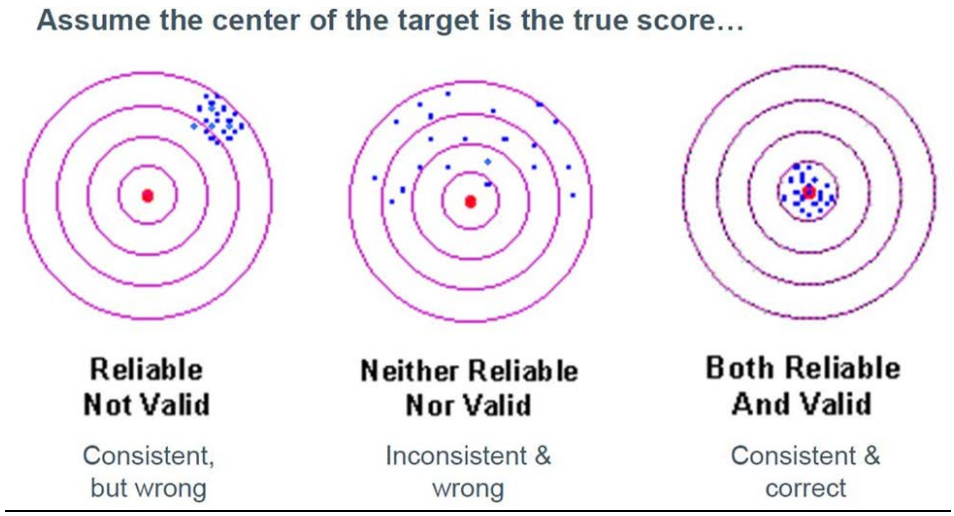
Algorithm #1 has been developer to assist Committee members and others to apply the criteria (Appendix A).

1b. Performance Gap, including disparities (must-pass) addresses the question of whether there is a quality problem that can be addressed by measurement. Data should demonstrate that there is an opportunity for improvement such as overall poor performance, substantial variation across providers, or variation for subpopulations (disparities in care).

> *Endorsed measures attaining high level performance*: For very strong measures that meet all other criteria, NQF may grant an "Inactive Endorsement with Reserve Status" when an endorsed measure fails to meet this criterion.

1c. For composite measures: quality construct and rationale (must-pass) 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. This subcriterion addresses whether there is a coherent quality construct and rationale that guided construction of the composite.

**2. Scientific acceptability of measure properties** (must-pass) reflects the extent to which the measure, as specified, produces consistent and credible results about the quality of care. The focus of this criterion is measurement science—not clinical science (which is the focus of the evidence subcriterion.) Measures that are reliable and valid enable users to make correct conclusions about the quality of care.




Measure developers conduct empirical analyses—collectively referred to as *measure testing*—in order to demonstrate the reliability and validity of a measure. Various methods and statistics can be used to quantify reliability and validity, although some may be more appropriate than others.

2a. Reliability (must-pass) considers the chance error (or "noise") in a measure result. All measures have some error—but when there is a lot of error in a measure, it can be difficult to know whether (or how much) variation in performance scores between providers is due to "real" differences between providers or to measurement error.

> *Specifications:* precise specifications with all codes, definitions and instructions to assure standardized calculation of the measure results
>
> *Empirical reliability testing:* Testing at the data element level addresses the *repeatability/ reproducibility* of the patient-level data used in the measure. Testing at the performance measure score level addresses the *precision* of the measure.

Algorithm #2 has been developer to assist Committee members and others to apply the criteria (Appendix A).

2b. Validity (must-pass) refers to the extent to which one can draw accurate conclusions about a particular attribute based on the results of that measure. In the context of quality performance measurement, a valid measure will allow one to make correct conclusions about the quality of care (i.e., a higher score on a quality measure reflects higher quality of care).

> *Empirical testing of validity*: Testing at the data element level typically addresses the correctness of the patient-level data elements used in the measure compared to an authoritative source. Score-level testing should link the concept of interest (that is being measured) to some other concept(s) via some a hypothesis about the relationship between them, then empirically investigate whether that hypothesis holds true.
>
> *Face validity*: The subjective determination that, on the face of it, a measure appears to reflect quality of care—is the weakest demonstration of validity, but is accepted by NQF for the validity criterion (if systematically assessed.)
>
> *Potential threats to validity*: Potential threats to validity that should be considered include whether: exclusions are justified, risk-adjustment should be applied and if so, is appropriate, meaningful differences in performance can be identified, results are comparable if different data sources/methods are used, and if missing data are an issue and are handled appropriately.

Algorithm #3 has been developed to assist Committee members and others to apply the criteria (Appendix A).

2d. For composite measures: empirical analysis supporting composite construction (must-pass) While subcriterion 1d addresses the conceptual basis of the composite performance measure, this subcriterion allows developers to demonstrate—via *empirical* analyses—that the choices made regarding which components are included in the composite performance score and how those components are combined actually fit with their concept of quality.

**3. Feasibility** (not must-pass) reflects the extent to which the data required to compute a measure are readily available and retrievable without undue burden, as well as the ease of implementation for performance measurement. The first two subcriteria under Feasibility relate to the burden of data collection, and the third subcriterion relates to ease of implementation.

3a. Required data elements routinely generated and used during care delivery

3b. Availability in electronic health records or other electronic sources OR a credible, near-term path to electronic collection is specified. For eMeasures, a summary of a feasibility assessment is required.

3c. Data collection strategy can be implemented

**4. Usability and Use** (not must-pass) reflects the expectation that endorsed measures not only will be used, but also ultimately will lead to improved patient outcomes. This criterion considers 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 to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency - ideally the measure should be used in at least one accountability application by the time of endorsement maintenance and be publicly reported within six years of initial endorsement.

4b. Improvement – Data on use of the measure should demonstrate improvement in quality.

4c. The benefits to patients outweigh evidence of unintended negative consequences to patients – feedback from the field is sought to better understand the use of the measure.

4d. Vetting by those being measured or others – This is a new sub-criterion for usability and use in 2016. It is not a must-pass criterion. It will be used to consider whether the measure is eligible for the "Endorsement+" designation.

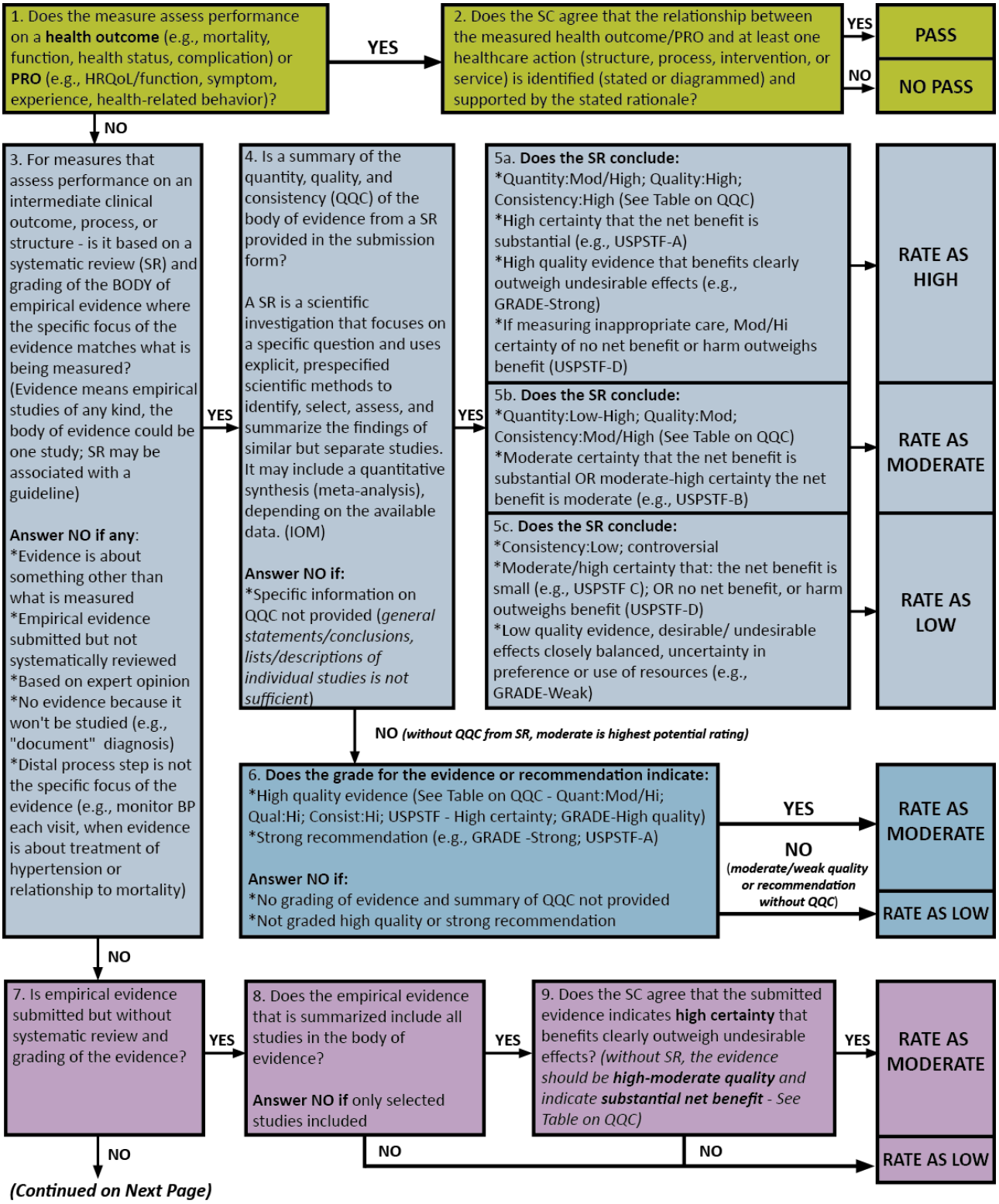
**5. Comparison to Related or Competing Measures** reflects the awareness that duplicative measures and/or those with similar but not identical specifications may increase data collection burden and/or create confusion or inaccuracy in interpreting performance results for those who implement and use those measures.

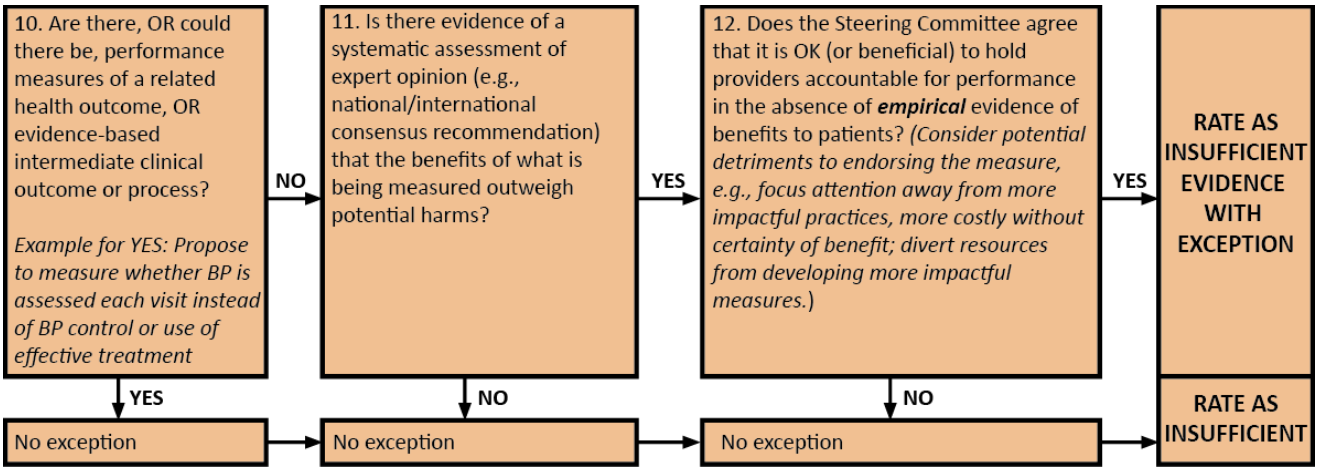
5a. Measure specifications are harmonized OR differences are justified - Harmonization of related measures should be done to the extent possible; differences in specifications should be justified.

5b. Superior measure is identified OR multiple measures are justified - The endorsement of multiple competing measures should be by exception, with adequate justification.

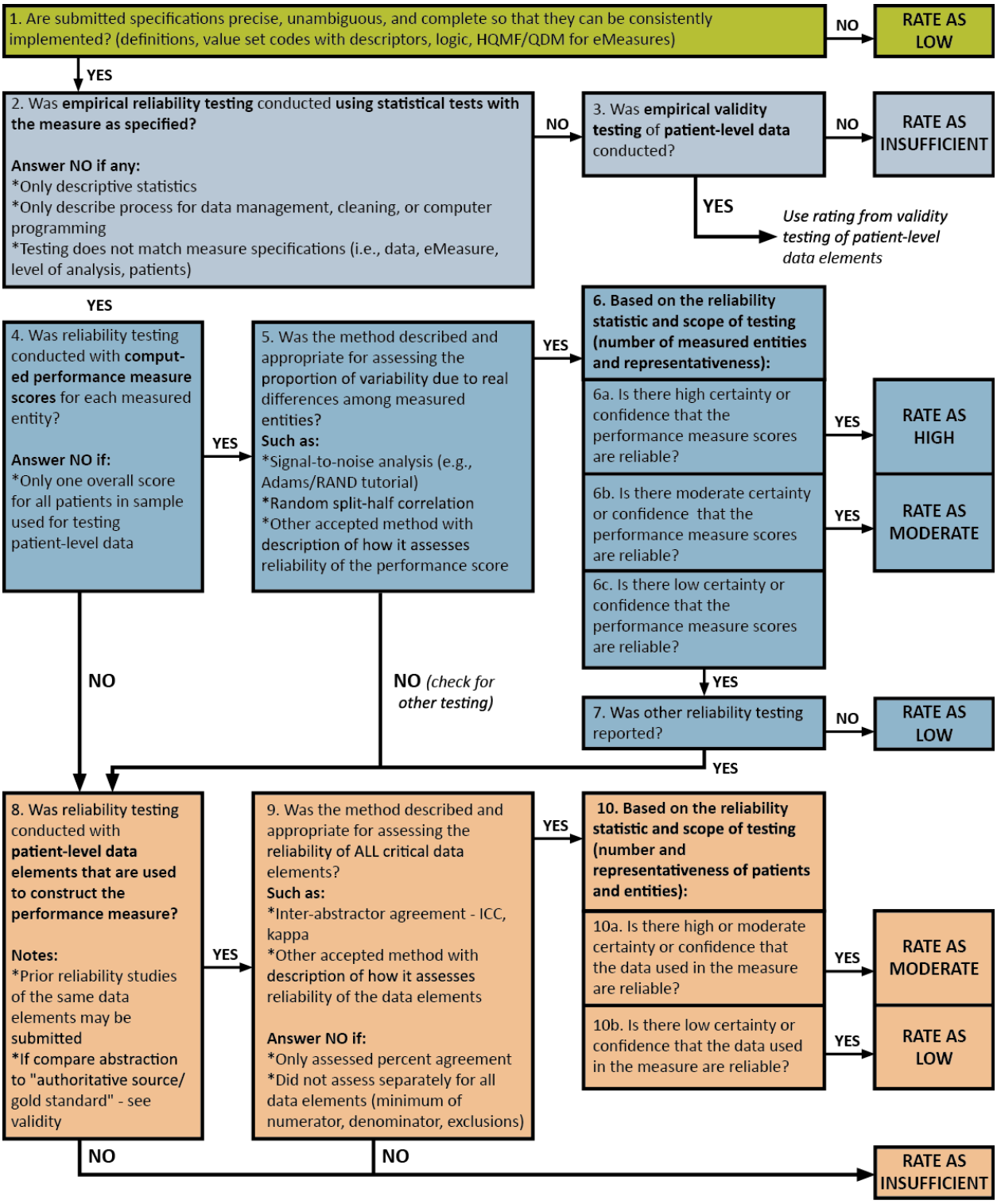
# APPENDIX

### Algorithm #1. Guidance for Evaluating the Clinical Evidence

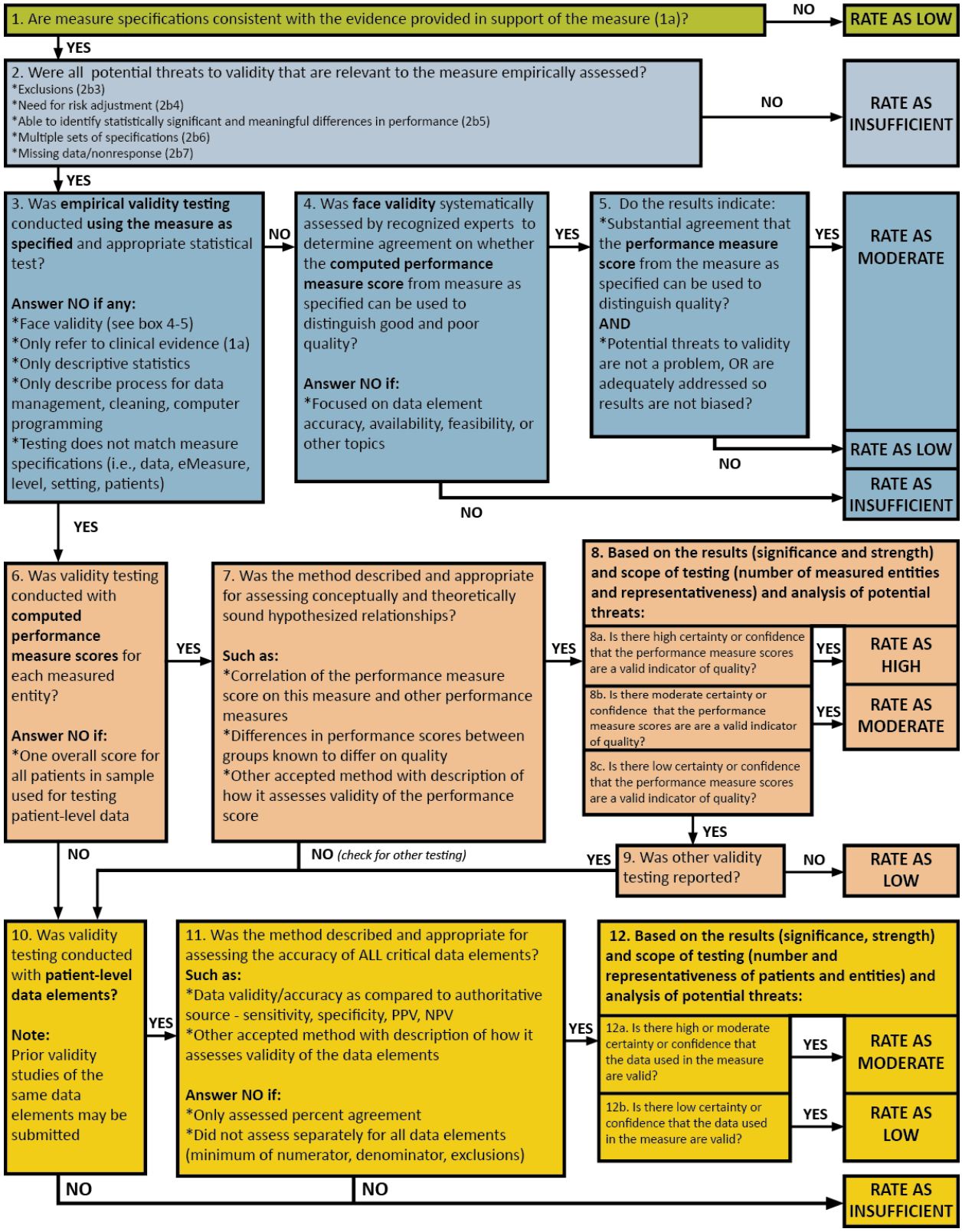



10. Are there, OR could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcome or process?
Example for YES: Propose to measure whether BP is assessed each visit instead of BP control or use of effective treatment
NO
11. Is there evidence of a systematic 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

## Algorithm #2. Guidance for Evaluating Reliability



## Algorithm #3. Guidance for Evaluating Validity



# NQF
# Measure Evaluation Criteria Overview

# NQF Measure Evaluation Criteria for Endorsement

**NQF endorses measures for accountability applications (public reporting, payment programs, accreditation, etc.) as well as quality improvement**.

- Standardized evaluation criteria
- CSAC oversees CDP process and 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

*Hierarchy and Rationale*

- **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**: Goal is to use for decisions related to accountability and improvement; if not useful, probably do not care if feasible
- Comparison to related or competing measures

# Criterion #1: Importance to Measure and Report

1. **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.

   **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

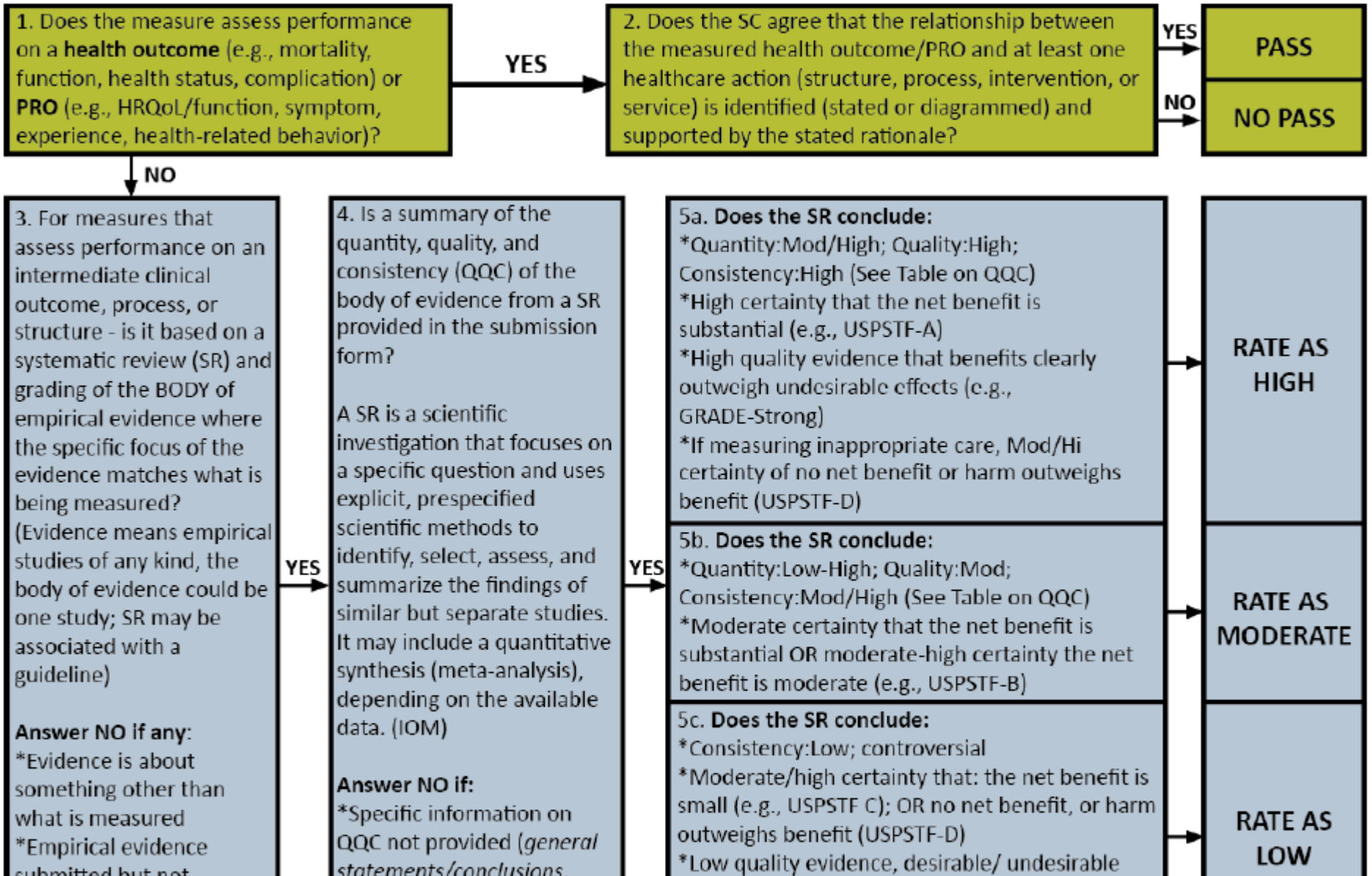
   **1c. Quality construct and rationale** (composite measures only)

# Subcriteron 1a:  Evidence

- Outcome measures
  - A rationale (which often includes evidence) for how the outcome is influenced by healthcare processes or structures.
- 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
    - » Empiric studies  (expert opinion is not evidence)
    - » Systematic review and grading of evidence
      - *Clinical Practice Guidelines – variable in approach to evidence review*

# Rating Evidence: Algorithm #1 – page 38

## Algorithm #1. Guidance for Evaluating the Clinical Evidence



# 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 health care delivery**

2a. **Reliability (must-pass)**

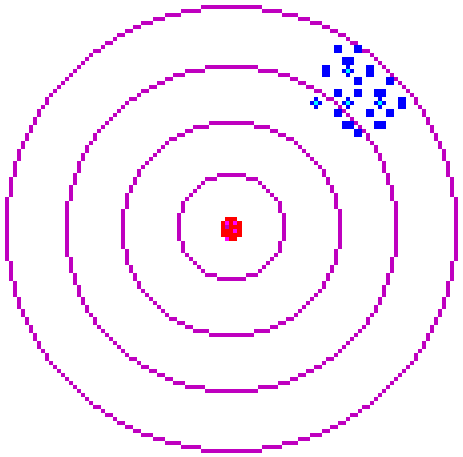
- 2a1. Precise specifications including exclusions
- 2a2. Reliability testing—data elements or measure score

2b. **Validity (must-pass)**

- 2b1. Specifications consistent with evidence
- 2b2. Validity testing—data elements or measure score
- 2b3. Justification of exclusions—relates to evidence
- 2b4. Risk adjustment—typically for outcome/cost/resource use
- 2b5. Identification of differences in performance
- 2b6. Comparability of data sources/methods
- 2b7. Missing data

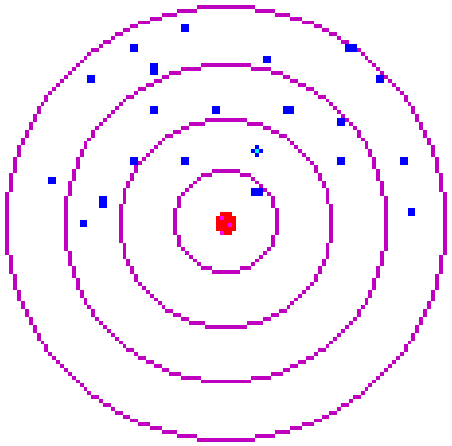
# Reliability and Validity

**Assume the center of the target is the true score…**



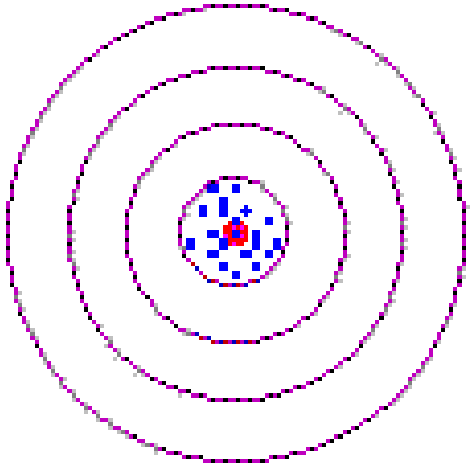
**Reliable Not Valid**

Consistent, but wrong



**Neither Reliable Nor Valid**

Inconsistent & wrong



**Both Reliable And Valid**

Consistent & correct

# Measure Testing

**Empirical analysis** to demonstrate the reliability and validity of the *measure as specified,* including analysis of issues that pose threats to the validity of conclusions about quality of care such as exclusions, risk adjustment/stratification for outcome and resource use measures, methods to identify differences in performance, and comparability of data sources/methods.

# Reliability Testing

- Reliability of the ***measure score*** refers to the proportion of variation in the performance scores due to systematic differences across the measured entities in relation to random variation or noise (i.e., the precision of the measure).
  - Example - Statistical analysis of sources of variation in performance measure scores (signal-to-noise analysis)

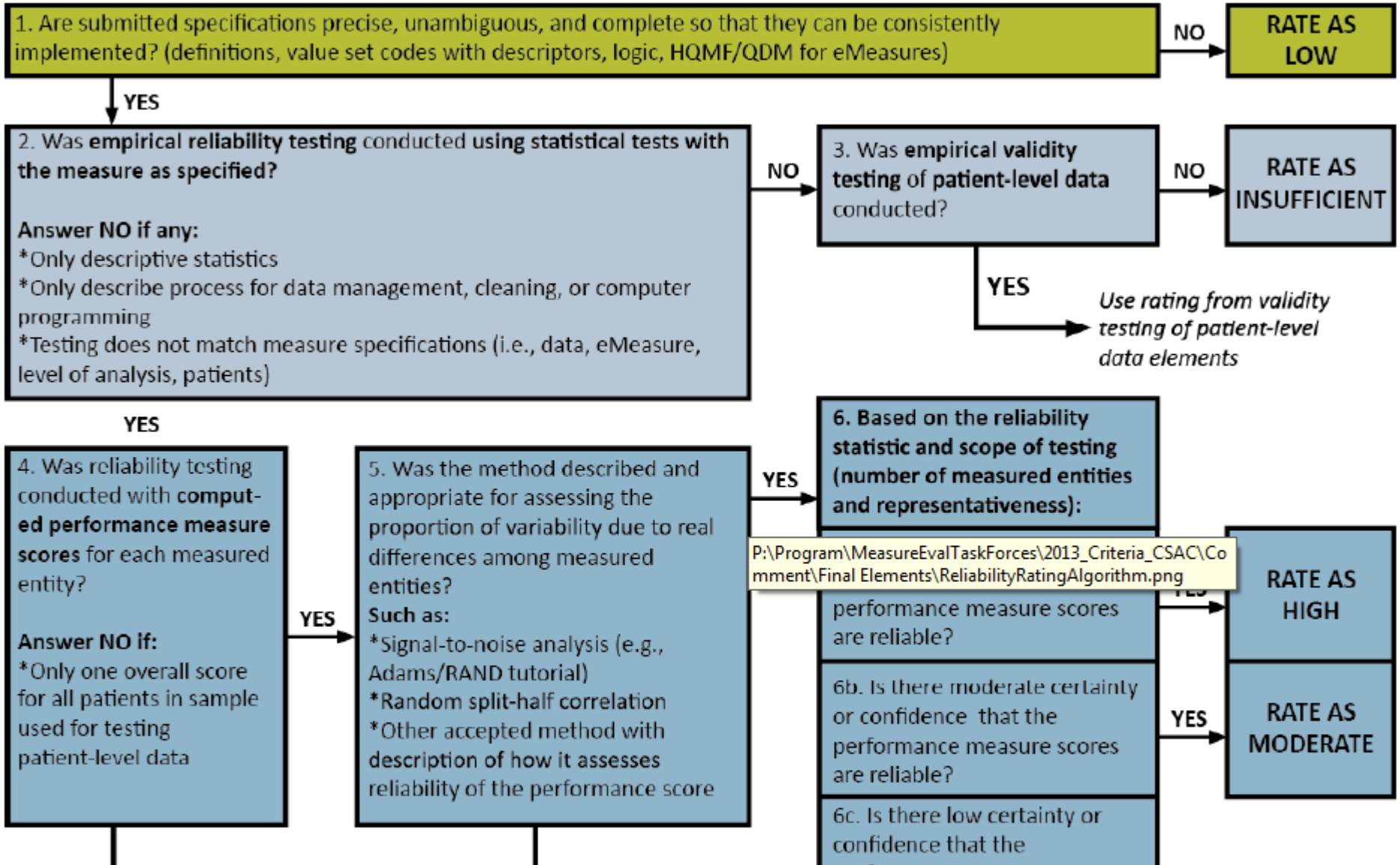
- Reliability of the ***data elements*** refers to the repeatability/reproducibility of the data and uses patient-level data
  - Example –inter-rater reliability

- Consider whether testing used an appropriate method and included adequate representation of providers and patients and whether results are within acceptable norms

- Algorithm #2

# Rating Reliability: Algorithm #2

## Algorithm #2. Guidance for Evaluating Reliability

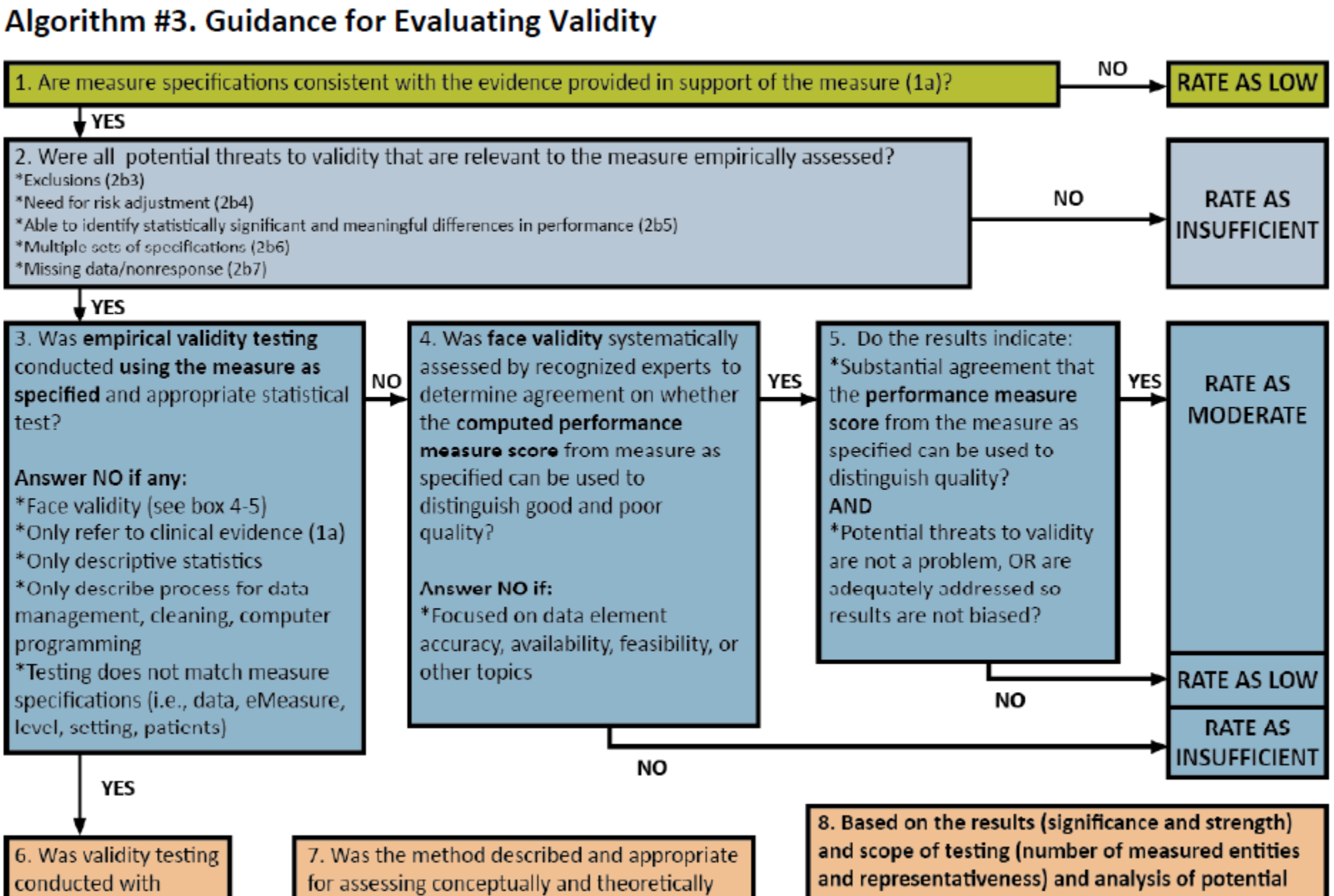



P:\Program\MeasureEvalTaskForces\2013_Criteria_CSAC\Comment\Final Elements\ReliabilityRatingAlgorithm.png

# Validity testing

- **Empirical testing**
  - *Measure score* – assesses a hypothesized relationship of the measure results to some other concept; assesses the correctness of conclusions about quality
  - *Data element* – assesses the correctness of the data elements compared to a "gold standard"
- **Face validity**
  - Subjective determination by experts that the measure appears to reflect quality of care

# Rating Validity: Algorithm #3



# Threats to Validity

- Conceptual
  - Measure focus is not a relevant outcome of healthcare or not strongly linked to a relevant outcome
- Unreliability
  - Generally, an unreliable measure cannot be valid
- Patients inappropriately excluded from measurement
- Differences in patient mix for outcome and resource use measures
- Measure scores that are generated with multiple data sources/methods
- Systematic missing or “incorrect” data (unintentional or intentional)

# 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**

# Criterion #4: Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

**4a: 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

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

**4c: 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).

**4d. Vetting by those being measured and others** - new sub-criterion for usability and use in 2016. It is not a must-pass criterion. It will be used to consider whether the measure is eligible for the "Endorsement+" designation.

# 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.