Key points for evaluating Scientific Acceptability

- NQF is not prescriptive about how empirical measure testing is done; similarly, NQF has 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). If multiple levels of analysis or data sources are specified, testing must be conducted for each level of analysis and each data source. If, for example, two levels of analysis are specified, but testing is conducted for only one, fully evaluate the testing that was done (don't just stop the evaluation)—however, the overall rating must be INSUFFICIENT. In this example, it is possible that the measure could be endorsed for the one level of analysis, but not for both.
- For most types of measures (but not all), NQF allows 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, consider whether testing used an appropriate method, included adequate representation of providers and patients, and whether results are within acceptable norms.

Key points for evaluating Reliability

- Precise specifications provide the foundation for achieving consistency in measurement. All data elements must be clearly defined.

Data element reliability: Addresses the repeatability/reproducibility of the data used in the measure

- Required for all critical data elements (i.e., those needed to calculate the measure score), or, at a minimum, for the numerator, denominator, and exclusions
- Not required if ***data element validity*** is demonstrated

Performance measure score reliability: Addresses the precision of the measure; indicates ability to ability to distinguish differences between providers that are due to quality of care rather than to chance.

- Common method is signal-to-noise analysis; have allowed point estimates and confidence intervals, if shown for all providers

Other considerations:

- For instrument-based measures (including PRO-PMs), reliability should be demonstrated for both the instrument and for the computed performance score.
- For composite performance measures, reliability should be demonstrated for the computed performance score.
- For measures that use ICD-10 coding: For Fall 2017 and CY2018 submissions, testing based on ICD-9 coding will suffice, but would prefer testing based on ICD-10 if available. For CY2019 and beyond, reliability testing should be based on ICD-10 coded data.
- For eMeasures (eCQMs): Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid. Reliability testing is not required if based on data from structured data fields.

Key points for evaluating Validity

- Validity refers to the correctness of measurement: that the measure is, in fact, measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided.
- Ratings for validity take into account testing information as well as consideration of potential threats to validity.

Data Element Validity: Analyzes agreement with authoritative source of the same information

- Required for all critical data elements (i.e., those needed to calculate the measure score), or, at a minimum, for the numerator, denominator, and exclusions
- Prefer sensitivity/specificity and positive (and negative) predictive values; have accepted kappa statistics, but not percent agreement only

Performance Measure Score Validity:

- We expect assessment of a hypothesized relationship of the measure results to some other concept, but labels (e.g., concurrent, predictive, etc.) don't really matter. Consider the merits of the hypothesized relationship, the method used to assess the relationship, and the results of the assessment.
- Ideally, multiple validation studies will be done over time, but this is not required.
- For new measures, we allow face validity (subjective determination by experts that the measure appears to reflect quality of care, done through a systematic and transparent process, that explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality, with degree of consensus and any areas of disagreement provided/discussed). For maintenance measures, empirical testing is expected; however, face validity may be accepted if you accept the justification provided by the developer for why empirical testing was not conducted.

Potential threats to validity include:

- Exclusions: must be supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure
- Risk adjustment approach through statistical modeling or risk stratification. This should be considered for all outcome/cost/resource use measures. If a measure is not risk-adjusted, this must be justified conceptually and/or empirically. Inclusion of social risk factors should be considered conceptually, and if there is a conceptual rationale, should be assessed empirically.
- Ability to identify differences in performance
- Comparability of data sources/methods
- Ensure that missing data do not bias results and/or that missing data are handled in a way that minimized bias

Other Considerations:

- For instrument-based measures (including PRO-PMs), validity should be demonstrated for both the instrument and for the computed performance score.
- For composite performance measures, validity should be demonstrated for the computed performance score. For new measures, a systematic assessment of content or face validity of the composite performance measure or empirical validation of the components will suffice.
- For measures that use ICD-10 coding (see Guidance for Measures Using ICD-10 Coding): Beginning with Fall 2017 submissions, updated validity testing must be submitted:
    - Submit updated empirical validity testing on the ICD-10 specified measure, if available
    - OR face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of quality
    - OR face validity of the ICD-10 coding scheme plus score-level empirical validity testing based on ICD-9 coding
    - OR face validity of the ICD-10 coding scheme plus data element level validity testing based on ICD-9 coding, with face validity of the measure score as an indicator of quality due at annual update
    - For CY2019 and beyond, validity testing should be based on ICD-10 coded data; if providing face validity, both face validity of the ICD-10 coding scheme plus face validity of the measure score as an indicator of quality is required.
- For eMeasures (eCQMs): Beginning September 30, 2017, all respecified measure submissions for use in federal programs (previously known as "legacy" eMeasures) will be required to conform to the same evaluation criteria as respecified measures – the "BONNIE testing only" option will no longer meet endorsement criteria.

Guidance from NQF's Scientific Methods Panel (SMP)

Guidance that has been formally incorporated into NQF's *Measure Evaluation Criteria and Guidance*

Reliability

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

Validity

If presenting score-level validation (typically via construct validity or known-groups analysis) the following should be included:

- Narrative describing the hypothesized relationships
- Narrative describing why examining these relationships (e.g., correlating measures) would validate the measure
- Expected direction of the association
- Expected strength of the association
- Specific statistical tests used (more detail is better)
- Results of the analysis
- Interpretation of those results (including how they related to the hypothesis and whether they have helped to validate the measure)

Requirements and Guidance that have not yet been formally incorporated into NQF's *Measure Evaluation Criteria and Guidance*

NOTE that these requirements and guidance ***may or may not be*** formally adopted by NQF. They are provided to give interested stakeholders an early indication of potential future changes.

- If conducting score-level reliability testing using the "split-half" approach, developers should assess via the ICC statistic, rather than via a Pearson or Spearman correlation.
- For all measure types, for both reliability and validity, NQF should require testing at both the data element level and the measure score level.