NATIONAL QUALITY FORUM
Driving measurable health improvements together

DRAFT Acceptable Reliability Thresholds (Version 3.2)

Measure developers have requested that the National Quality Forum (NQF) provide additional support and clarification for reliability testing thresholds based on the varying testing levels, methods, and interpretation standards available for testing. Reliability is one of the two "must-pass" criteria for scientific acceptability in NQF's Measure Evaluation Criteria[i]. Although a single reliability threshold is not possible, several evidence-based guidelines and empirical analysis options are available to assess reliability that are routinely used by developers for the two levels of analysis: 1) person-/encounter-level reliability testing (i.e., data element testing) and 2) accountable reporting entity-level reliability testing (i.e., performance or measure score testing).

Based on recent Scientific Methods Panel (SMP)[ii] activity, the NQF staff drafted a framework to provide more clarity for acceptable thresholds of reliability. This revised *draft* is a framework for the SMP members to continue discussing the components, content, and context for recommending objective, empirical, and science-based reliability testing thresholds for initial endorsement, maintenance, and implementation purposes during the July 29, 2021 SMP advisory web meeting. This tool is also intended to aid measure stewards and developers in conducting reliability testing in measure submissions, and NQF committees to evaluate person-/encounter-level and accountable reporting entity-level testing.

The thresholds identified here are expressed as point estimates. Measure developers are strongly encouraged to present point estimates for reliability, as well as confidence intervals around those estimates, sensitivity analyses to show how reliability of the measure will depend on sample size of the entities being evaluated, and one or more examples of potential rates of misclassification in specific use scenarios (e.g., if the measure is going to be used to divide entities into quintiles for a form of "star rating").

Person-/Encounter-Level Reliability Testing (i.e., data element testing)

Approach *(Test)*	Purpose	Range	Threshold
Internal consistency *(e.g., Cronbach's Alpha)*	The internally consistency of items in a multi-item scale.	0 to 1	0.7
Inter-rater agreement *e.g., (Cohen's Kappa)*	The inter-rater agreement of qualitative items correcting for chance.	−1 to +1	0.4
Test-Retest Reliability *(Intraclass coefficient [ICC] or Pearson correlation)*	Extent to which two measurements of the same concept at different times agree.	-1 to +1	0.5
Linear Relationships *(e.g., Pearson correlation coefficient)*	Agreement between two measures of the same concept.	−1 to +1	0.6

Accountable Reporting Entity Level Reliability Testing (i.e., performance measure score testing)

Approach *(Test)*	Testing Purpose	Range	Threshold
Signal to Noise Ratio (SNR) or Inter-Unit Reliability (IUR)	The precision attributed to an actual construct versus random variation.	−1 to +1	0.5
Split-half reliability *(Intraclass coefficient, with correction for full sample with Spearman-Brown formula)*	Agreement between two measures of the same concept derived from split samples drawn from the same entity at a single point in time.	0 to 1	0.5

References (To be added for each)

[i] https://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx

[ii] https://www.qualityforum.org/Measuring_Performance/Scientific_Methods_Panel.aspx