NATIONAL QUALITY FORUM
Driving measurable health improvements together

Draft Acceptable Reliability Thresholds

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, thresholds may differ. Although determining a single reliability threshold is not possible, several evidence-based guidelines and empirical analysis options are available to assess reliability. These 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 measure score testing). Reliability is one of the two "must-pass" criterion for scientific acceptability in NQF's Measure Evaluation Criteria.

Based on recent Scientific Methods Panel (SMP) activity, NQF staff drafted a framework to provide more clarity for *unacceptable*, *adequate*, and *high* reliability thresholds of acceptability. This *draft* is a framework to discuss the components, content, and context for recommending objective, empirical, and science-based reliability testing thresholds for initial endorsement, maintenance, and implementation purposes during the May 4, 2021 SMP advisory web meeting. This tool is also intended to aid measure stewards and developers to conduct reliability testing for measure submissions, and NQF standing committees to evaluate reliability testing.

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

	Testing Purpose	Range	U	A	H
Inter-rater Agreement ***(Cohen's Kappa Coefficient)***	The inter-rater agreement of qualitative items correcting for chance.	0 to 1	< 0.7	0.7 ≥ 0.9	≥ 0.9
Internal Consistency ***(Cronbach's Alpha)***	The internal consistency of items in a multi-item scale.	−1 to +1	< 0.6	0.6 ≥ 0.9	≥ 0.9
Test-Retest ***(Intraclass Coefficient [ICC])***	Descriptive statistics used when quantitative findings are units and organized into ranks or groups.	0 to 1	< 0.6	0.6 ≥ 0.9	≥ 0.9
Signal to Noise (SNR) ***(Pearson Correlation Coefficient)***	The precision attributed to an actual construct versus random variation.	−1 to +1	< 0.6	0.6 ≥ 0.9	≥ 0.9

Note: Reliability threshold definitions: U = Unacceptable, A = Adequate, and H = High

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

	Testing Purpose	Range	U	A	H
Signal to Noise (SNR) ***(Intraclass Coefficient)***	Descriptive statistics used when quantitative findings are units and organized into ranks or groups.	0 to 1	< 0.5	0.5 ≥ 0.9	≥ 0.9
Test/Retest ***(Simple Correlation)***	The relationship or association between dependent and independent variables.	0 to 1	< 0.5	0.5 ≥ 0.9	≥ 0.9
Test/Retest ***Intraclass Correlation Coefficients [ICC])***	Descriptive statistics used when quantitative findings are made on units and organized into ranks or groups.	0 to 1	< 0.5	0.5 ≥ 0.9	≥ 0.9
Various Split-half Methods ***(Spearman-Brown Coefficient)***	The consistency of scores when a test is repeated on a population.	−1 to +1	< 0.4	0.4 ≥ 0.8	≥ 0.8

Note: Reliability threshold definitions: U = Unacceptable, A = Adequate, and H = High

Other Reliability Considerations

	Testing Purpose	Range	U	A	H
Interunit Reliability (IUR)	The proportion of variation that comes from inter-provider differences.	0 to 1			
Profile Interunit Reliability (PIUR)	The proportion of variation that comes from inter-provider differences with extreme outcomes.	0 to 1			
Classification Stability	The limits of stability between and within test groups.	0 to 1			

Note: Reliability threshold definitions: U = Unacceptable, A = Adequate, and H = High