Partnership for Quality Measurement
Powered by Battelle

National Consensus Development and Strategic Planning for Health Care Quality Measurement

Reliability Guidance for the Endorsement and Maintenance (E&M) of Clinical Quality Measures

Prepared by:
Battelle
505 King Avenue, Columbus, Ohio 43201
September 2025

The analyses upon which this publication is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services.

Table of Contents

List of Tables

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted: Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Introduction

Overview

Health care quality measurement is essential for evaluating and enhancing the performance of health care providers and systems in the United States, as well as ensuring that patients receive safe, effective, and patient-centered care.

The Partnership for Quality Measurement (PQM), formed by Battelle, a Centers for Medicare & Medicaid Services (CMS)-certified consensus-based entity (CBE), plays a crucial role in shaping the health care quality measurement landscape. Under PQM's measure endorsement and maintenance (E&M) process, measures submitted for endorsement consideration are evaluated for scientific acceptability—including reliability.

For each level of reliability testing conducted (i.e., person/encounter level or accountable entity level), measure developers must describe their methodology, provide results, and interpret how results support an inference of reliability. The Blueprint Measure Lifecyle content on the Measures Management System (MMS) Hub provides general guidance about the types, uses, and tests for measuring reliability.

This document builds on that guidance and provides measure developers with the tools to effectively apply reliability methods and interpret results in preparation for endorsement consideration, ensuring their measures are robust and scientifically sound.

Purpose and Objectives

This document is intended to guide developers and stewards by supporting their ability to:

- Review the various tests of reliability and statistical methods at both the person/encounter level and the accountable entity level;
- Determine the best method for assessing reliability in their quality measure; and
- Evaluate mitigation strategies and tradeoffs associated with common challenges in assessing reliability.

Intended Audience

This resource guide is intended to support measure developers and stewards in the calculation of person- or encounter-level (data element) and accountable entity-level (performance score) reliability within the context of health care clinical quality and cost/resource use measurement. Measure developers and stewards can also use this guide in preparation for measure submission to PQM for E&M and Pre-Rulemaking Measure Review (PRMR) processes.

In addition, other interested parties (e.g., PQM committee members, quality improvement teams, patient advocates) may use this guide to better understand reliability methodologies and to support CBE evaluation of measures.

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted: Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Reliability in the Context of Quality Measurement

Clinical quality and cost/resource use measures are evaluated at two levels: the person/encounter level and the accountable entity level.

Person or Encounter Level

At the person or encounter level, reliability is generally defined as the absence of ambiguity in the measure specification and the repeatability of the data collection process for elements such as diagnoses, procedures, or lab results recorded across different encounters and individuals. High reliability at this level ensures that the underlying data used to construct the measure are consistent and reproducible. While electronically captured data elements (e.g., from structured fields within electronic clinical quality measures or administrative claims data) are often presumed to be more reliable than those obtained through manual abstraction—due to increased standardization, reduced human subjectivity, and repeatable data entry processes—reliability ultimately depends on the clarity of data definitions, consistency of capture across systems and users, and alignment with clinical workflows.

Accountable Entity Level

At the accountable entity level—such as providers, hospitals, or health systems—reliability assesses whether observed differences in performance reflect true quality rather than random variation. High reliability at this level ensures that comparisons across entities are based on reproducible results, not on random fluctuations or inconsistent data.

Without sufficient reliability at both levels, quality measurement and accountability efforts risk being misleading or unfair. Random inaccuracies in data recording or sampling error can distort performance results, undermining the validity of comparisons between providers or organizations.

Table 1 summarizes the key terms and definitions.

Reliability

Reliability is the degree to which a measure repeatedly and consistently produces the same result.

ISO/IEC 25020 — Quality measurement framework (2019)[1]

Table 1. Key Terms and Definitions

Term	Definition
Consistency	The degree of variation in the process or construct a measure is intended to reflect.

[1] International Organization for Standardization. (2019). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality measurement framework (ISO/IEC Standard No. 25020:2019).

Term	Definition
Reliability	The degree to which a measure repeatedly and consistently produces the same result.[1] A characteristic of the measure when applied to an entity in a population of entities. It represents one type of competing cause of variation—random error or chance—among others affecting the measure focus (e.g., process or outcome).
Repeatability	The extent of variation inherent in a single method (i.e., a data collection method).[1]
Reproducibility	The amount of variation due to other sources (i.e., other than quality).[1]
Stability	The degree of variation over time (in either the performance score or the process the measure is intended to reflect).
Validity	An overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy of appropriate interpretations and actions on the basis of [the measure].[2,3]

Reliability and Stability

Stability is impacted by both reliability and consistency. A measure that is not reliable may also not be stable due to chance. However, it is possible (and often common) for a measure to be reliable but not stable. When possible, developers are encouraged to track entity-level performance scores longitudinally to assess the measure's stability over time (i.e., the change in performance scores over time).

Reliability and Validity

Reliability is a component of the validity argument. Specifically, it helps to rule out chance as a competing cause of variation, alongside other sources such as confounders and counter-acting mechanisms. A measure that is not reliable results in a validity judgement with considerable residual risk.

Transparency

Multiple statistical methods are used to estimate reliability, and these can yield different results, especially in settings with small sample sizes. The lack of standardization in reliability estimation methods can affect the fairness and comparability of provider performance assessments, as differences in reliability estimates may reflect methodological artifacts rather than true differences in quality. Transparent documentation and justification of the chosen reliability estimation approach are essential to support the integrity of the measure and accountability efforts.

---

[2] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
[3] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Types of Data

To appropriately select statistical tests for reliability, it is important to consider the type of data collected at the person/encounter level and how the performance score (accountable entity level) is constructed (i.e., observed rate, risk-adjusted rate, composite). Different types of data require different analytic approaches.

Categorical data are qualitative and consist of distinct groups or labels. These data can be either nominal, where categories have no inherent order (e.g., blood type, gender), or ordinal, where categories have a meaningful order, but the intervals between them may not be uniform (e.g., education level, satisfaction ratings). Categorical data can also be classified by the number of response categories:

- Binary (dichotomous) data include only two possible categories (e.g., yes/no, pass/fail). Many quality measures are binomial as they may assess the presence or absence of a condition or event.
- Multinomial (polytomous) data include three or more categories (e.g., eye color, marital status). Ordinal data are categorical and represent ordered categories.

Numerical data are quantitative and may be either:

- Discrete, meaning they represent countable whole numbers (e.g., number of children, test scores), or
- Continuous, meaning they can take on any value within a specified range and are measurable (e.g., height, weight, temperature).

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted:

Tests of Reliability

Measure developers can assess reliability using a variety of tests for both the person/encounter level and the accountable entity level. For each test, we describe common methods, appropriate use cases (based on data type or performance score), methodological variations, examples, and CBE-recommended thresholds.

Person or Encounter Level

Test-Retest Reliability

Test-retest reliability evaluates the consistency of results when the same individual completes the same instrument or measure on two separate occasions. This assessment is appropriate when there is a rationale for expecting stability, rather than change, over a short to medium time interval. For example, a measure developer may administer a patient satisfaction survey to the same group of patients 2 weeks apart to test if responses are consistent. Test-retest reliability is not suitable for repeated measurement of disease symptoms or intermediate outcomes that are expected to improve or deteriorate over time (e.g., blood pressure control). A coefficient of stability is calculated to quantify the strength of the relationship between the results obtained at the two time points.

Methods: Depending on the data type, the methods for test-retest reliability include:

- Spearman's Rank Correlation (SR)[4] is for ordinal data. Spearman's Rank Correlation is a non-parametric method, meaning it does not require the data to follow any particular probability distribution including a normal distribution or require a linear relationship between results at the two time points. It is especially useful when working with ranked or skewed data and can be applied to any type of measure with patient-level data.
- Pearson Correlation Coefficient[5] is used for continuous data that are normally distributed and when a linear relationship is expected.

[4] Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

[5] Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242. https://doi.org/10.1098/rspl.1895.0041

- Intraclass Correlation Coefficient (ICC) can also be used for continuous or ordinal data and assesses reliability by examining the consistency of measurements, often by randomly splitting patient-level data within each entity and comparing reliability across subsets ("split-half").

These methods produce values on a 0-1 scale, and a higher correlation value indicates greater test-retest reliability.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.5

Inter- and Intra-Rater Agreement

Inter-rater agreement assesses the extent to which two or more individuals provide consistent judgments or ratings when evaluating the same information. This assessment is commonly applied in quality measures that require manual review or data abstraction to evaluate consistency among abstractors. For instruments, inter-rater agreement is generally not used with standard respondent-completed surveys, as each participant reports their own experiences. However, it may be applicable when a trained observer or interviewer is responsible for rating behaviors, responses, or performance based on a survey or checklist.

Intra-rater agreement measures the consistency of scores assigned by the same individual across two or more occasions. This assessment is particularly relevant in quality measurement when manual review or human judgment is involved in data collection. For example, it applies when a single individual (e.g., abstractor or coder) is expected to make consistent decisions if they review the same chart at different times.

Methods: The same methods can be used to assess both inter-rater and intra-rater reliability, depending on the type of data:

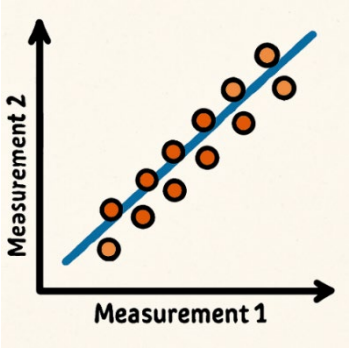
- Cohen's Kappa[6] is appropriate for categorical data (e.g., yes/no responses) and adjusts for agreement that could occur by chance. A value of 1.0 indicates perfect agreement, while a value of 0 indicates agreement equivalent to chance.
- Gwet's AC1, like Cohen's Kappa, is a chance-corrected agreement statistic, but it is particularly well-suited for binary or categorical data that may be imbalanced (e.g., when the prevalence of a category is very high or very low).

[6] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

- Kendall's Tau is a non-parametric correlation coefficient used with ordinal (ranked) data. It is less sensitive to small sample sizes and is useful for assessing the consistency of rankings over time or between raters.
- Percent agreement is the number of ratings that agree divided by the total number of ratings. While easy to compute, it does not adjust for agreement that may occur by chance and may therefore overestimate reliability. As such, it is most appropriate as a basic or preliminary measure of agreement.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.4[7]

Linear Relationship



Linear relationship examines whether person- or encounter-level data from repeated measurements track together in a predictable, linear way.

Methods:

- The Pearson Correlation Coefficient measures the strength and direction of a linear relationship between two continuous variables. However, it can be misleading in reliability analyses because it assesses association rather than agreement.

CBE Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.6

Instrument-Derived Measures: Construct Reliability

Internal Consistency

Internal consistency refers to the degree to which items correlate with one another and reliably capture the same underlying construct. Often used to evaluate an instrument, such as a survey, internal consistency reflects how well the questions collectively measure a single concept (e.g., patient satisfaction or provider communication).

Methods: Commonly used methods for internal consistency include:

- Cronbach's alpha (α) for continuous or ordinal data (Cronbach, 1951)[8]

[7] The 0.4 threshold is based on the moderate agreement range for the kappa statistic, as described by Landis and Koch (1977).
[8] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

- Kuder-Richardson Formula 20 (KR-20)[9] for binary (dichotomous) items.

Both methods produce values on a 0-1 scale, in which higher values indicate greater internal consistency.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.7

Table 2 summarizes the appropriate methods for each test of reliability at the person or encounter level. Table 3 includes example use cases.

Table 2. Summary of Reliability Testing Approach Methods for Person/Encounter-Level Data

Method	Type of Data	Internal Consistency[10]	Test-Retest	Inter-rater Agreement	Intra-rater Agreement	Linear Relationship
Cronbach's alpha	Continuous or ordinal	X				
KR-20	Binary	X				
ICC	Continuous		X	X	X	X
Spearman's Rank Correlation	Ordinal (non-parametric)		X			
Pearson Correlation Coefficient	Continuous (assumes linearity)		X			X
Cohen's Kappa	Categorical			X	X	
Gwet's AC1	Binary or categorical (robust to imbalance)			X	X	
Kendall's Tau	Ordinal (non-parametric, small n)			X	X	
Percent Agreement	Categorical			X	X	

[9] Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

[10] Often used to evaluate the reliability of an instrument.

Table 3. Examples of Reliability Testing Approaches for Person/Encounter-Level Data

Measure Data Type	Measure of Reliability	Measure of Reliability Tests	Example Use Case
Survey/instrument – continuous data (e.g., scale)	Internal consistency[10]	Cronbach's alpha	A scale instrument developed for a patient-reported outcome measure assessing pain levels should undergo internal consistency reliability testing. This ensures it consistently measures the construct of pain intensity. Achieving a Cronbach's alpha > 0.7 is crucial for credibility, and refining scale items may be necessary. This process supports accurate patient assessments and treatment evaluations in clinical settings.*
Survey/instrument – single items or continuous data (e.g., scales)	Test-retest reliability	Correlation coefficient (e.g., Pearson or Spearman); Percent agreement	The same survey items are administered to the same respondents at two different points in time, with the interval long enough to prevent recall bias but short enough to avoid real change. This assesses the stability of responses over time. High test-retest reliability indicates that the instrument yields consistent results when the underlying construct has not changed, supporting the credibility and interpretability of the measure. For example, a patient-reported outcome measure for pain is given to patients twice, 1 week apart, to ensure that responses are stable if their pain has not changed.
Administrative claims data (e.g., ICD codes, procedure codes)	Inter-rater agreement (between coders)	Cohen's kappa; Percent Agreement	Two coders assign diagnosis codes to patient encounters. Reliability is tested to ensure that codes are applied consistently across coders and encounters.
Retrospective chart abstraction (including registry data abstracted from medical records)	Inter-rater agreement (between abstractors)	Cohen's kappa; Percent Agreement	Two abstractors independently review patient charts to determine whether a pain assessment was documented during a hospital stay. Inter-rater reliability is calculated to ensure that the abstraction process yields consistent results. High inter-rater agreement indicates that the data abstraction process is well-defined and reproducible, which is essential for credible quality measurement. Consistent abstraction reduces the risk of random error and increases confidence that differences in pain documentation reflect true differences in care, not measurement error.
Any type of patient-level data (e.g., continuous, categorical, scale)	Split-half reliability	ICC	Within each hospital or clinic, patient-level pain scores are randomly split into two groups. The average pain score is calculated for each half, and the ICC is computed to assess the consistency of the measure

Measure Data Type	Measure of Reliability	Measure of Reliability Tests	Example Use Case
			across the two groups.

*When constructing a performance measure using patient scores derived from an established instrument or scale, the internal consistency reliability of that scale serves as robust evidence for data element reliability. If the reliability of these instruments has typically been rigorously tested and documented, this existing validation can be leveraged or cited to support claims of instrument reliability.

Accountable Entity Level

Measure developers should conduct accountable entity-level reliability testing with performance scores for each measured entity. It is important to note that reliability is a characteristic of the measure applied to an entity in a population of entities. Table 4 includes example use cases.

Signal-to-Noise

Signal-to-noise is used to estimate how much of the overall variation in entity-level performance scores is due to systematic differences in performance between entities (the "signal") as opposed to random variation or measurement error ("the noise"). Generally, signal-to-noise is preferred for binomial measures. A higher value indicates greater confidence that observed differences reflect actual entity performance rather than chance, thereby supporting more meaningful comparisons across providers.

Methods: Two methods to test signal-to-noise include:

- Adams (2009),[11] which estimates signal-to-noise by using the beta-binomial parameter estimates for each entity.
- Nieser and Harris (2024),[12] which estimates signal-to-noise based on the sample size and the beta-binomial parameter estimates. We recommend that developers use the Nieser Harris approach (given Adams results in a reliability of 1.0 when the performance score is 0% or 100%).

Guidance

For binomial measures (i.e., observed rate), we recommend using the signal-to-noise method calculated as suggested by Nieser and Harris (2024).

Both methods assume the measure follows a beta-binomial distribution—meaning the probability of occurrence can vary from group to group—and can only be applied to binomial data.

[11] Adams, J. L. (2009). *The reliability of provider profiling: A tutorial* (TR-653). RAND Corporation. https://www.rand.org/pubs/technical_reports/TR653.html

[12] Nieser, K. J., & Harris, H. S. (2024). Comparing methods for assessing the reliability of health care quality measures. *Statistics in Medicine, 43*(23).

Variations:

- Interunit reliability (IUR) is a variation of the signal-to-noise method that is focused on between-entity reliability.[13] The IUR uses total and pooled between-entity variances. A higher IUR means more of the observed variation is due to systematic differences between providers (stronger signal). A lower IUR means more of the observed variation is due to noise. Measure developers can apply the Spearman-Brown (SB) prophecy formula[14] to obtain an entity-level estimate:

$$\hat{r}_i = \frac{k_i \hat{r}}{1 + (k_i - 1)\hat{r}}$$

  where $\hat{r}$ is the overall reliability estimate, $\hat{r}_i$ is the entity-level reliability estimate, and k_i is the ratio of the entity-level sample size to a measure of central tendency (such as mean or median) of the sample size across all entities.

- Empirical Bayes[15] is another form of estimating signal-to-noise. This method can be applied to binomial data only and assumes that performance scores follow normal distribution. This approach "borrows strength" from entities with larger patient or case volume to improve estimates of reliability for entities with smaller patient or case volume. Therefore, it is particularly useful when measure developers anticipate small sample sizes.

- Logistic regression is a commonly used method for estimating reliability, especially for risk-adjusted measures with binomial data. In this approach, a hierarchical (multi-level) logistic regression model is fit, with the entity (e.g., hospital, provider) included as a random effect. This allows for estimation of both between-entity variance (signal) and within-entity variance (noise). The reliability estimate is then calculated as the ratio of between-entity variance to total variance, similar to the IUR.

While shrinkage (a statistical technique used in methods such as Empirical Bayes and hierarchical logistic regression to "borrow strength" for low-volume entities) can reduce noise and improve the reliability of estimates, it may also obscure the true performance of individual entities by pulling their performance scores toward the average, which can mask truly high or low performers—especially when sample sizes are small. To evaluate whether shrinkage is applied in a fair and neutral manner, measure developers should examine which entities are being adjusted and the magnitude of those adjustments. It is important to assess whether shrinkage is balanced—affecting both high and low performers proportionally—or whether certain types of entities are disproportionately impacted. Developers should also be aware of

[13] He, K., Dahlerus, C., Xia, L., Li, Y., & Kalbfleisch, J. D. (2020). The profile inter-unit reliability. *Biometrics*, *76*(2), 654–663. https://doi.org/10.1111/biom.13167
[14] Kingston, N., & Tiemann, G. (2010). Spearman–brown prophecy formula. In Encyclopedia of research design (Vol. 0, pp. 1403-1404). SAGE Publications, Inc., https://doi.org/10.4135/9781412961288.n427
[15] Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, *78*(381), 47–55.

problematic volume-outcome relationships related to small entitles (indicating a propensity to be worse performers). In those instances, the shrinkage is not "balanced" and is misleading. To mitigate this issue, developers should shrink to a volume-specific mean, rather than the overall mean. Without transparency into these adjustments, determining the validity and fairness of results produced by shrinkage-based methods is difficult to determine.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.6 for 70% or more of the accountable entities

Split-Half (ICC)

The ICC measures correlation at the patient or encounter level. To assess accountable entity level reliability, a formula must be applied to obtain an entity-level distribution of performance scores. Measure developers can apply the SB prophecy formula to a single ICC value to obtain an entity-level estimate of the ICC based on the entity's sample size. ICC is best suited for measures with continuous data (e.g., excess days in acute care [EDAC]).

Methods:

- Recent evidence[16] suggests that averaging a series[17] of resampled split-half reliability estimates provides a more accurate picture of reliability than performing the split-half method just once. This is done using bootstrap resampling, where the data are repeatedly split and sampled—with replacement (meaning each patient can be selected more than once)—and the results are averaged. This approach is widely used and helps to improve the repeatability of reliability estimates. In each resampling round, two samples are created by randomly selecting patients with replacement. For each entity, reliability is calculated by averaging the ICCs across all resamples. However, by sampling with replacement, the bootstrap can over- or under-represent certain observations, potentially distorting the original data. This method works best when data are independent and similarly distributed.
- Permutation sampling, is similar to bootstrap resampling but involves selecting multiple split samples without replacement. Permutation sampling is less likely to introduce bias from over- or under-representing certain observations, as it preserves the original sample sizes and relationships within the data. This approach is particularly well-suited for complex data that are not independent and identically distributed, such as patients nested within a specific provider. However, permutation sampling can be complex and computationally intensive, and the number of unique permutations may be limited in smaller datasets. The reliability estimate for each entity is calculated as the average ICC, with the SB formula applied across all permutations.

---

[16] Nieser, K. J., & Harris, A. H. S. (2024). Split-sample reliability estimation in health care quality measurement: Once is not enough. *Health services research*, *59*(4), e14310. https://doi.org/10.1111/1475-6773.14310

[17] Developers should assess the stability of bootstrap estimates rather than rely solely on a fixed number (e.g., 100). A useful visual check is to plot the cumulative mean of the estimates; if it stabilizes as more samples are added, the number of bootstraps is likely sufficient.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.6 for 70% or more of the accountable entities

Spearman Rank

The Spearman-Brown formula can also be applied to a Spearman rank correlation coefficient to calculate entity-level reliability based on the size of each entity. Spearman rank correlation is best for measures based on ordinal data that are not continuous or normally distributed (e.g., a survey measures that uses a 3-point Likert scale of poor, fair, good).

Guidance

For continuous measures (i.e., risk-adjusted rate, ratio, hierarchical model, composite), we recommend applying the SB formula to either ICC or SR correlation coefficient and permutation sampling.

Methods: Like ICC, developers can also use bootstrap resampling or permutation resampling to improve reliability estimates for Spearman rank correlation coefficients. In bootstrap resampling, the reliability estimate for each entity is calculated as the average Spearman rank correlation. With permutation resampling, the reliability estimate for each entity is the average Spearman rank correlation, with the Spearman-Brown formula applied across all permutations.

CBE-Recommended Value for Endorsement (i.e., Acceptable Threshold): ≥ 0.6 for 70% or more of the accountable entities

Table 4. Examples of Reliability Testing Approaches for the Accountable Entity-Level Data

Type of Performance Score	Measure of Reliability Tests	Example Use Case
Binomial (i.e., observed rate)	Signal-to-noise calculated as suggested by Nieser and Harris (2024)	As part of improving paint management, hospitals are being evaluated on the percentage (i.e., rate) of inpatients who had pain assessment documented during their stay. To determine if the measure can reliably distinguish rates between hospitals, use the signal-to-noise method, as suggested by Nieser and Harris, to estimate the reliability of this measure across hospitals. This involves calculating the signal (the true variation in performance across hospitals) and the noise (random variation due to sampling error within each hospital).
Binomial (i.e., observed rate) – small sample size/case volume	Signal-to-noise test and applying Empirical Bayes reliability estimation	In situations where a hospital has a small patient volume, the observed rate (such as the percentage of inpatients with pain assessment documented) can be highly influenced by random variation. To address this, the Empirical Bayes method is used. First, the observed rate is calculated for each hospital, and the true variation between hospitals (signal) and the random variation within hospitals (noise) are estimated. The Empirical Bayes approach then adjusts each hospital's observed rate by "shrinking" it toward the overall average rate, with the amount of shrinkage determined by

Type of Performance Score	Measure of Reliability Tests	Example Use Case
		the hospital's sample size and the estimated variance between hospitals. The shrunken estimate has lower prediction error (compared to the unobserved "true" score). Hospitals with smaller sample sizes are adjusted more, which helps reduce the impact of random noise and stabilize the estimates. Finally, the reliability of the measure for each hospital is calculated as the ratio of signal to total variance (signal plus noise), indicating the extent to which the observed differences reflect true differences in quality rather than chance. This method ensures more accurate and fair comparisons across hospitals, regardless of their patient volumes.
Continuous (i.e., risk-adjusted rate, ratio, hierarchical model, composite)	Intraclass correlation coefficient or Spearman rank correlation, both with Spearman-Brown prophecy formula adjustment and permutation sampling	Hospital pain management is being evaluated using a risk-adjusted pain management composite score, consisting of components such as pain assessment, timely intervention, and patient-reported pain relief. To test if this composite reliably differentiates hospital performance, for each hospital, calculate a risk-adjusted mean pain score or. For each hospital within the health system, the hospital is incorporating a composite pain score. Randomly split patient data within each hospital, calculate the measure for each half, and use ICC or SR (with permutation and Spearman-Brown adjustment) to estimate reliability.

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted: Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Challenges in Assessing Reliability

Measure developers may encounter challenges in accessing accountable entity-level reliability due to data limitations or testing samples. When a developer is unable to meet the CBE accountable-entity reliability threshold of 0.6, they should explore potential mitigation strategies and carefully weigh the tradeoffs that may result from implementing such strategies. Measure developers should aim to reduce harm caused by 1) using a low reliability measure, 2) low participation, and/or 3) unbalanced shrinkage. In their submission, measure developers should explain the underlying issues for low reliability and justify any decisions related to imposing mitigation strategies and their tradeoffs. Below, we include a table outlining common challenges, mitigation strategies, associated tradeoffs, and guiding questions to consider when evaluating measures that apply these approaches.

Table 5. Reliability Challenges, Mitigation Strategies, and Tradeoffs

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Small Sample Size/Case Volume	Aggregate data across multiple years to increase sample size	Slower feedback to accountable entities; data may be outdated and less actionable	• Are measure results timely enough for the intended purpose? • Are measure results consistent across years, suggesting slower feedback may be acceptable? • Do users prefer greater stability over rapid feedback?
	Combine individual measures into a composite	Loss of detail for individual measure areas; may obscure poor performance in specific areas	• Does the composite maintain conceptual coherence? • Is the sum better than the individual components? • Are individual component results available for review? • Are there processes to address poor performance in specific areas?
Some Low-Volume Entities	Use hierarchical modeling to "shrink" extreme values toward the mean	Adds complexity; may mask the performance of outliers (very high or low performers)	• Is there transparency in how entities are adjusted? • Do the reliability estimates of high and low performers change over subsequent measurement periods? • Does this approach preserve meaningful performance differences? • Are users comfortable interpreting adjusted performance scores?
	Report at a higher level to combine data across entities (e.g., system level)	Loss of granularity; may weaken local accountability	• Will users still find results meaningful and actionable? • Can entities review their results internally for more specific insights?

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted: Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
			• Is local accountability still possible with aggregated reporting?
	Exclude low-volume providers by setting a minimum threshold	Excludes small providers; may omit safety-net or underserved populations	• Is a low reliability measure better than no measure at all due to its importance and no alternative measure? • What percentage and types of entities are excluded? • Are excluded providers serving vulnerable populations? • Are alternative monitoring or quality improvement strategies in place (e.g., confidential feedback) for excluded entities?
Infrequent Events or Rare Outcomes	Combine similar outcomes or events into one measure	Loss of specificity; may conflate clinically distinct issues	• Does combining events still reflect meaningful performance? • Are stakeholders aligned on what is appropriate to group? • Can disaggregated data still be analyzed for root cause analysis?
	Extend measurement period (e.g., over multiple years)	Delays event detection; potential increased risk from slower signal recognition	• Is delayed feedback acceptable for this event type? • Are risks of delayed detection outweighed by improved reliability? • Can other mechanisms provide early warning signals?
Little Variation Between Entities	Increase measure sensitivity (e.g., use more granular categories or continuous scoring)	Increased complexity may reduce interpretability	• Will users understand and accept more sensitive but complex measures? • Are there subpopulations where variation is present?
Measure is Topped Out	Determine if the performance standard or threshold should be raised	Demotivates entities if new targets are perceived as unattainable	• Are there higher standards or new evidence that justify a more challenging measure?

When is Low Reliability Acceptable?

Low reliability does not necessarily mean a measure is not useful; its utility depends on available alternatives and the consequences of its use. When the measure focus is closer to the continuum of "should not happen," (e.g., never events) less emphasis may be placed on reliability since chance does not need to be ruled-out as a cause. Reasoning should rule out other explanations for the association between the response to the quality program and the measure focus. This means that even if a measure has low reliability, it can still be useful if the reasoning behind its use effectively rules out alternative explanations for the observed outcomes. Essentially, the measure's utility is not solely dependent on its reliability but also on the strength of the reasoning that supports its use in the given context.

There are circumstances where low reliability may be acceptable (see Table 6) and others where it is not warranted due to the potential for harm or waste (see Table 7). In either case, measure developers should provide a clear justification for accepting or mitigating lower reliability, based on the specific context and use of the measure. As a best practice, the submission should describe the decision-making process, including any input received from interested parties (e.g., technical expert panels, public comment) that informed the final determination.

Table 6. Scenarios Where Low Reliability May Be Acceptable

Scenario in Which Low Reliability May Be Acceptable	Description	Example
Direct Causal Relationship	The measure focuses on an outcome with a clear, direct causal link to the intervention, with no alternative mechanisms or confounding person-level factors.	A measure of surgical site infections following a specific procedure, where the procedure itself is the only plausible cause.
No Alternative for Decision-Making	When no other reliable measures are available, using a low reliability measure may still provide valuable information for decision-making.	A rural health program uses a low reliability measure of patient satisfaction because no other patient-reported data is collected in the area.
Availability of a Complementary High-Reliability Measure	A related structure, process, or outcome measure with high reliability is available to support interpretation and use of the lower reliability measure.	A measure of complication rates (low reliability) is paired with a measure of adherence to surgical protocols (high reliability).
Mitigation Would Reduce Validity or Participation	Strategies to improve reliability (e.g., combining data over multiple years) would reduce the measure's validity or discourage provider engagement.	Extending data collection to 3 years would increase reliability but obscure recent improvements and disincentivize participation in an accountability application.

Table 7. Scenarios Where Low Reliability May Not Be Acceptable

Scenario in Which Low Reliability Is Not Acceptable	Description	Example
Disproportionate Impact on Vulnerable Populations	When specific, identifiable populations are more likely to receive care from providers with low reliability scores, leading to potential biases.	Low reliability in accountable entities (e.g., safety-net hospitals) serving marginalized communities may result in unfair penalties that disproportionately affect underserved populations.
Potential for Material Harm	When decisions informed by the measure could result in serious consequences for providers or patients, such as public reporting, payment adjustments, or loss of trust.	A low reliability mortality rate measure used in a payment program could lead to unjust financial penalties for accountable entities without accurately reflecting performance.
Risk of Waste, Futility, or Injustice	When low reliability compromises the measure's overall value (e.g., by reducing its importance, validity, or usability), leading to wasted resources or unjust outcomes.	A low reliability measure of provider communication is used to rank clinicians, despite weak correlation with actual care quality, resulting in misleading comparisons and reputational harm.

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services.

E&M Testing Requirements

Reliability testing requirements vary depending on whether a measure is being submitted for initial endorsement or maintenance. The minimum requirements for each are outlined below.

When submitting reliability testing information, measure developers must clearly specify the level of testing (i.e., person or encounter level vs. accountable entity level), the method(s) used, statistical results for each test, and a plain-language interpretation of the results to aid understanding.

For example, a signal-to-noise reliability score of 0.658 indicates that 65.8% of the variation in performance scores across entities reflects true differences in performance, rather than random error. In general, the higher the score, the more reliable the measure.

Initial Endorsement

When submitting a new measure for initial endorsement, developers must demonstrate reliability at the person/encounter level for all critical data elements (i.e., those used to calculate the numerator, denominator, and any exclusions). Reliability testing is required for data elements that are manually collected or extracted using natural language processing (NLP), such as those abstracted from medical records or clinical notes. However, reliability testing is not required for data elements captured electronically through structured fields (e.g., in eCQMs) or from administrative claims, provided that prior evidence (e.g., prior empirical research, literature) of reliability exists for those data elements and data types.

For NLP-based data collection, developers should report reliability statistics from at least two sites using a standardized template of common person-level features (e.g., age, diagnosis, medication status) to ensure consistency and comparability. Any discrepancies identified during testing must be clearly explained. For example, differences in inter-rater agreement may be due to variation in documentation practices across sites.

Developers may also reference existing evidence, such as published literature, prior empirical analyses, and/or existing data auditing protocols to support the reliability of critical data elements and reduce the need for new testing. This evidence should employ the testing approaches identified in Tables 2 and 3, and developers should clearly identify whether the existing evidence includes all critical data elements.

Maintenance Endorsement

Measure developers must submit empirical testing at the accountable entity level for maintenance measures. The associated thresholds for accountable entity-level reliability testing are applied to the accountable entity (e.g., facility, clinician, health plan), not the mean or median across entities. To evaluate the distribution of reliability scores across accountable entities, measure developers must complete Accountable Entity-Level Reliability Testing Results by Denominator, Target Population Size within the Full Measure Submission (FMS) form (Table 8). Stepwise instructions for completing the reliability table are included below.

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted:

Table 8. Accountable Entity-Level Reliability Testing Results (Full Measure Submission Form Table 2)

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability													
Mean Performance Score													
N of Entities													
N of Persons / Encounters / Episodes													

Step-by-Step Guide for Calculating Accountable Entity-Level Reliability Testing Table

1. **Data Collection:**
   - Gather the overall mean, minimum, and maximum performance scores.
   - Collect the count of measured entities and the number of persons/encounters/episodes for each entity.
2. **Organize Entities into Deciles:**
   - Sort entities based on the number of persons/encounters/episodes (denominator).
   - Divide the sorted list into 10 equal parts (deciles), where Decile 1 contains entities with the smallest number of persons/encounters/episodes, and Decile 10 contains entities with the largest number. In order for 10 equal parts to be reflected, developers may add a small random number to the denominator (e.g., 1/1000000).
3. **Calculate Mean Reliability and Performance Score for Each Decile:**
   - For each decile, calculate the mean reliability by averaging the reliability scores of entities within that decile.
   - Calculate the mean performance score by averaging the performance scores of entities within that decile.
4. **Count Entities and Persons/Encounters/Episodes for Each Decile:**
   - Count the total number of entities within each decile.
   - Sum the number of persons/encounters/episodes for all entities within each decile to get the total for that decile.
5. **Determine Minimum and Maximum Reliability:**
   - For minimum reliability, identify the entity with the smallest number of persons/encounters/episodes within each decile and note its reliability value.
   - For maximum reliability, identify the entity with the largest number of persons/encounters/episodes within each decile and note its reliability value.
6. **Compile Results:**
   - Create a summary table or report that includes:
     - Mean reliability for each decile.

  - Mean performance score for each decile.
  - Total number of entities in each decile.
  - Total number of persons/encounters/episodes in each decile.
  - Minimum reliability value for the entity with the smallest n in each decile.
  - Maximum reliability value for the entity with the largest n in each decile.

7. **Review and Validate:**
   - Ensure all calculations are accurate and data is correctly organized.
   - Validate the results by cross-checking with the original data to ensure consistency.

By following these steps, developers can systematically organize and analyze the performance scores and reliability values across different deciles.

Guidance Authors

Battelle

Anna Michie, MHS, PMP, Senior Measurement Scientist

Stephanie Peak, PhD, Social Scientist

Matthew Pickering, PharmD, Principal Measurement Scientist

Jeffrey Geppert, JD, EdM, Senior Research Leader

Laura Aume, MS, Data Scientist

William White, MS, Data Scientist

 The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services.


Partnership for
Quality Measurement
Powered by Battelle