Journal of Integrated Knowledge
Vol. 2 No. 11 (2026)

Reliability of a Single Judge and Reliability of a Panel

Gabriel Baker[1] · Simon Barnes[2] · Jason Reed[3]

[1] Uzbekistan · [2] Moldova · [3] Brunei

Published: November 05, 2026 | Article 1023

Abstract

A panel average can be more stable than one judge's score while remaining sensitive to panel composition. Reliability must be attached to the assessment configuration it describes.

Reliability of a Single Judge and Reliability of a Panel

Abstract

A panel average can be more stable than one judge's score while remaining sensitive to panel composition. Reliability must be attached to the assessment configuration it describes.

Introduction

The data-driven judging topic of Zhao et al. (2022) permits summaries for a single assessor and for an assembled panel. Those summaries are not interchangeable. Combining assessors can reduce some idiosyncratic variation, but shared interpretation or common bias may persist. A review should identify whether the claim concerns the panel actually observed or another panel drawn from a wider assessor population. Adding more ratings from the same configuration does not automatically establish the latter. The reliability account becomes useful when the number, composition, and intended use of assessors remain attached to the score being evaluated.

Different scales need different agreement questions

Different scales indeed require different agreement questions, as evidenced by the distinction between nominal-scale and categorical-agreement perspectives. For instance, Cohen's (1960) approach is suitable for classification problems involving discrete categories, while Landis and Koch’s (1977) perspective is more relevant for situations where judgments are made on a continuous or ordinal scale. The key point here is that reliability measures should be tailored to the specific nature of the ratings or criteria being assessed. A narrow set of entries might not capture enough variability to accurately gauge a judge's consistency, whereas a broader range could highlight differences among judges and entries. This underscores the importance of specifying the intended case distribution when reporting reliability coefficients, as it influences how these measures are interpreted. For example, if a study uses a specific subset of items that do not vary much in difficulty or relevance, the resulting reliability coefficient might overstate the judge's consistency. Conversely, using a wider range of items could reveal more nuanced differences among judges and entries, providing a more comprehensive assessment of reliability. Thus, it is crucial to clearly define the scope of the evaluation when discussing reliability coefficients to avoid misinterpretation.

Rating designs and agreement

In rating designs and agreement, the distinction between predicting and explaining is crucial for understanding reliability (Shmueli, 2010). When assessing a single judge's reliability versus that of a panel, the goal should guide the interpretation of the results. While a panel average might offer more stability due to varied compositions, it can also mask differences in individual judges' performance (Laird & Ware, 1982). The random-effects framework underscores the importance of aligning the model structure with the underlying processes that define groups (Laird & Ware, 1982). This is essential for ensuring that the reliability measure accurately reflects the intended assessment configuration and does not merely reflect the convenience of the software used (Bates et al., 2015).

The decision pathway deserves its own evaluation

The evaluation of the decision pathway deserves careful consideration, as it can significantly influence the practical outcome and reliability. A single judge's score might be more unstable and less sensitive to variations in panel composition compared to a panel average. The stability of a panel average is valuable because it remains robust even if different judges are involved. However, this stability comes at the cost of sensitivity, which means that the panel's overall score may not reflect the nuances of individual assessments. A future assessment should thus be comprehensive, encompassing the entire pathway from observation through interpretation to response. This approach allows for a nuanced understanding of the reliability and effectiveness of the decision-making process. For example, cases where further review or deferral is necessary are critical. Such instances highlight the necessity of additional analysis and do not necessarily resolve underlying issues. By evaluating the whole pathway, we can distinguish between a useful signal that aligns with the original scientific question and an effective decision process that might be flawed but still appears to work well due to its robustness.

Assignment and structured variation precede adjustment

In Laird and Ware (1982) and Bates et al. (2015), mixed-model perspectives acknowledge that different units, such as judges and entries, contribute varying levels of variation. When designing assessments, it is crucial to consider these variations and align them with the assignment design. If judges are assigned to evaluate different sets of entries, their scores can be influenced by both case difficulty and severity, leading to unadjusted averages that do not accurately reflect the reliability of individual judges or panels. To address this, a fitted adjustment should be made only as interpretable as the design and underlying assumptions support it.

Moreover, it is important to account for the uncertainty in assessments from the relevant independent units rather than treating each rating as a separate assessor. This approach helps maintain data volume distinct from the number of independent challenges to the proposed quality interpretation. By focusing on the assignment and structured variation before making adjustments, we ensure that the reliability measures are appropriately linked to the assessment configuration they describe.

Rubric structure and measurement

Rubric structure and measurement matter because they determine how consistent and reliable assessments are. In the context of reliability, the choice between a single judge's score and a panel average reflects different assessment configurations. A panel average can be more stable than a single judge's score while still being sensitive to the composition of the panel. For instance, if a panel of three judges is assessing the same set of items, their scores might vary more widely than a single judge's score. However, a panel average might be more consistent and thus more reliable. This stability is important because it ensures that the assessment results are not just a lucky coincidence but a true reflection of the item being rated.

The intraclass-correlation (Shrout & Fleiss, 1979) work highlights this distinction. It distinguishes reliability questions by the rating design and intended use. For example, a single-assessor score is different from a panel average. A single judge's score can be influenced by their own biases or mood on any given day, leading to variability. In contrast, a panel average reduces this variability because it averages over multiple judges. This difference underscores why rubric structure and measurement are crucial in ensuring that the reliability of an assessment is accurately described.

The choice between a single judge's score and a panel average is significant due to their different structures and intended uses. The stability and consistency provided by a panel average make it more

reliable than a single judge's score, especially when the goal is to reduce variability and ensure that the assessment results are not idiosyncratic.

Transport requires a named destination

A named destination for evaluation clarifies the context within which results are assessed, distinguishing them from other potential interpretations or applications. When a result lacks a clear endpoint, its generality is ambiguous. It may apply to one scenario but not another, depending on factors such as population, measurement conventions, or intended use. For instance, a judge's reliability might be robust in one context but less so in another with different criteria or participants. By defining the destination, we can more clearly delineate which aspects of a result are transferable and under what conditions they hold true. This approach allows for a balanced assessment: acknowledging that even if a panel's average is more stable than a single judge's score, this doesn't automatically mean it's more reliable across all contexts or uses. Thus, the destination provides a clear boundary for the successful application of reliability measures, preventing claims of generality from being too broad and encompassing every possible scenario.

Consensus should not manufacture a clean target

The section on consensus should not manufacture a clean target highlights the importance of maintaining the original differences and uncertainties within a panel's decisions, rather than treating them as definitive. When a panel resolves disagreements and issues a final decision, the initial variations and ambiguities are crucial for understanding the measurement process. By framing the consensus as unqualified truth, models that mimic the panel's decisions risk appearing independently validated when they may not fully account for the nuances and complexities of individual ratings.

To preserve this critical information, future records should document the original ratings, reasons for revision, and unresolved alternatives. This approach ensures that the final decision does not obscure the uncertainty that is essential for evaluating the reliability of judges or automated assessments. By doing so, we maintain a transparent record that allows for a more comprehensive analysis of the measurement process, rather than creating a sanitized view that may mislead about the true variability and complexity within the panel's decisions.

Statistical interpretation of assessor evidence

The ICC guideline underscores the importance of specifying the model, rating unit, and agreement definition when reporting reliability (Koo & Li, 2016). Failing to do so leaves readers uncertain about what aspect of assessment is being described. Coefficient alpha, while useful for assessing internal consistency among test items, does not directly address the reliability of judgments by different assessors (Cronbach, 1951). Thus, assessing a panel's average reliability requires separate evidence from evaluating the consistency of individual judges' scores.

The categorical-agreement analysis offers another perspective on interpreting observer agreement, highlighting that summary measures depend on the specific classification problem and its population (Landis & Koch, 1977). It emphasizes the distinction between disagreement about difficult cases and inconsistency in applying a settled rule. This qualitative distinction is crucial for accurately interpreting reliability findings, reinforcing the need to preserve these distinctions in reporting.

Agreement does not define a good rubric

Judges can agree on a rubric, yet their agreement may be superficial. If they share a clear rule but omit a property that motivated the contest, their agreement is likely due to shared understanding rather than a well-defined assessment target. Conversely, if judges disagree, it suggests the rubric allows for multiple interpretations, indicating a lack of a single, definitive measure. This highlights the importance of separating reliability from construct validity. Reliability measures how consistent scores are across different occasions or raters, whereas construct validity assesses whether the scores truly reflect the intended criteria. Agreement does not necessarily imply a good rubric. It could indicate that the rubric lacks clarity or has multiple valid interpretations. As Shmueli (2010) notes, judges can agree on a rubric because they share a clear rule, but this rule may omit properties that influenced the contest. A rubric's reliability and construct validity thus require distinct analyses, as agreement does not define a good rubric. Evaluating the instrument itself is essential to understand its strengths and weaknesses, rather than merely focusing on individual assessor performance.

Related criteria do not establish rater agreement

The internal-structure perspective, as discussed by Cronbach (1951) and Tavakol and Dennick (2011), focuses on the relationships among items rather than on whether judges consistently rate similar entries. This distinction is critical because it can obscure the actual assessment configuration being evaluated. For example, a rubric might contain closely related criteria but have assessors apply them differently, leading to high internal consistency without true agreement. Conversely, low agreement could indicate ambiguous wording or poorly defined criteria rather than an inherent poor judgment from the assessors. By separating these perspectives, we can diagnose more accurately and avoid overreliance on a single summary statistic like reliability.

A quality score has a development history

The quality score used for assessment must be distinct from the development history of the scale, as it may inadvertently reinforce conformity to the instrument used during training. This can lead to an evaluation that rewards consistency with the model rather than assessing its broader utility or construct validity. When separating feedback and observations, it is crucial to preserve all ratings and exclusions, maintaining a focus on the assessment process. This approach aligns with Rubin's perspective (1976), emphasizing the importance of sensitivity to criterion weights and multiple comparisons in identifying assessors. By doing so, we ensure that the quality score remains an independent measure of reliability, not a mere validation of the development scale itself.

Severity and inconsistency are not interchangeable

Severity and inconsistency are not interchangeable concepts, as they reflect different aspects of scoring reliability. A judge might consistently score higher than peers while maintaining the same relative rankings, indicating a systematic tendency rather than random variation. Conversely, another judge could have an average score but exhibit inconsistent decisions on specific entries. The Intraclass Correlation Coefficient (ICC) framework by Shrout and Fleiss (1979) and Koo and Li (2016) underscores the importance of distinguishing these patterns.

Absolute scores can be influenced by severity, meaning that higher scores might indicate more severe issues in certain cases. However, this does not necessarily mean that judges are making inconsistent decisions. Rather, it suggests that the criteria being assessed carry varying degrees of difficulty. On the other hand, an ordered decision may be more sensitive to these differences, as a consistent pattern across

multiple entries could indicate a reliable systematic tendency, while occasional discrepancies might reflect variations in the difficulty or importance of the tasks assigned. Therefore, assessing the consistency and severity of scoring patterns separately is crucial for understanding the reliability of the assessment configuration. This distinction allows for targeted feedback to judges, turning apparent inconsistency into a diagnostic tool rather than an unexplained variation among raters.

Conclusion

One consequence of this distinction is that reliability measures may not always reflect true agreement among judges. For instance, a panel's average score might be more stable than individual judges but still not accurately capture the diversity in how criteria are applied. This can lead to a misinterpretation of the panel's performance, as high reliability does not necessarily mean high consistency or agreement in judgment. An open question is how to further refine the assessment configuration to better distinguish between reliability and other aspects of judgment, such as severity or consistency, without losing the benefits of a panel's collective approach.

References

Koo, T. K., Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15(2), 155-163*. https://doi.org/10.1016/j.jcm.2016.02.012

RUBIN, D. B. (1976). Inference and missing data. *Biometrika, 63(3), 581-592*. https://doi.org/10.1093/biomet/63.3.581

Shrout, P. E., Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability.. *Psychological Bulletin, 86(2), 420-428*. https://doi.org/10.1037/0033-2909.86.2.420

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika, 16(3), 297-334*. https://doi.org/10.1007/BF02310555

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1)*. https://doi.org/10.18637/jss.v067.i01

Laird, N. M., Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics, 38(4), 963*. https://doi.org/10.2307/2529876

Landis, J. R., Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33(1), 159*. https://doi.org/10.2307/2529310

Tavakol, M., Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2, 53-55*. https://doi.org/10.5116/ijme.4dfb.8dfd

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(1), 289-300*. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20(1), 37-46*. https://doi.org/10.1177/001316446002000104

Shmueli, G. (2010). To Explain or to Predict?. *Statistical Science, 25(3), 289-310*. https://doi.org/10.1214/10-STS330

Zhao, R., Guo, Y., Ma, X., et al. (2022). Research on Evaluation System of the Judge Quality in Students' Contest Based on Data Driven. *2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM)*. doi:10.1109/ICEKIM55072.2022.00233.