Annals of Science and Inquiry
Vol. 2 No. 11 (2026)

Judge Consistency and Judge Quality Are Different Constructs

Peter Thompson[1] · Daniel Holmes[2] · Paul Lewis[3]

[1] Indonesia · [2] Martinique · [3] Christmas Island

Published: November 02, 2026 | Article 1018

Abstract

Judges can agree with one another while sharing a biased interpretation of quality. Reliability should therefore remain distinct from the validity of the contest judgment.

Judge Consistency and Judge Quality Are Different Constructs

Abstract

Judges can agree with one another while sharing a biased interpretation of quality. Reliability should therefore remain distinct from the validity of the contest judgment.

Introduction

The judge-quality topic of Zhao et al. (2022) raises two questions: how consistently an instrument is applied and whether it measures the intended quality. A panel could produce closely matching scores because all members rely on the same limited cue. Their agreement would describe consistency, not independently establish the appropriateness of that cue. A review should identify the external or conceptual basis for the quality judgment before treating consistency as its validation. Disagreement can also be meaningful when the rubric leaves legitimate interpretations open. The useful assessment keeps agreement, bias, and construct definition separately visible rather than combining them into one favorable reliability label.

Different scales need different agreement questions

In the context of distinguishing scales and their disagreement questions, the key distinction is between judge consistency and quality, which is crucial for understanding how scales are applied. Cohen's nominal-scale approach and Landis and Koch's categorical-agreement perspective, while effective for binary classifications, may not be adequate for continuous ratings or ordered criteria.

Moreover, the variability within the population of entries can significantly impact perceived consistency. A narrow selection might yield less score variation, making judges appear more consistent due to minimal differences among entries. Conversely, a broad set of entries could expose inconsistencies through larger differences. Neither approach is universally applicable. Thus, specifying the intended case distribution is essential. This ensures that an aggregate coefficient accurately reflects the variability within the specific context rather than being an isolated property detached from the judging criteria.

Rating designs and agreement

In the context of judges' ratings and agreement, it is crucial to distinguish between how judges' ratings are designed and the level of agreement among them. The distinction between explaining and predicting helps clarify that a model's predictive success does not automatically imply a deep understanding of the underlying mechanisms (Shmueli, 2010). Thus, while judges may agree on a rating scale, this does not necessarily mean they are interpreting quality in the same way. For instance, two judges might consistently rate contestants high because they both have a bias toward positive outcomes, rather than due to a shared understanding of what constitutes quality. This highlights the need for careful design of rating scales and agreement questions that probe the judges' actual interpretation of quality rather than their mere agreement on scores.

Moreover, the choice of rating designs should align with the scientific goals and the intended generalizability of the findings. A model's success in prediction does not equate to a valid explanation of the observed relationship (Bates et al., 2015). Therefore, judges' ratings should be designed with consideration for the specific context and the research questions at hand, rather than simply fitting a convenient model structure. For example, if the goal is to understand how judges' biases affect quality assessment, then rating scales should be crafted to reveal these biases explicitly, not just to produce consistent scores.

Finally, false-discovery-rate control in statistical comparisons remains a critical concern (Benjamini & Hochberg, 1995). While successful estimation of a model does not validate its specific interpretation or causal claims, it is essential to ensure that any favorable outcomes are not mistakenly interpreted as independent confirmations. In the context of judges' ratings, this means carefully examining whether agreement among judges is due to shared biases or substantive understanding of quality.

A useful quantity need not be a complete construct

An operational quantity, chosen for its utility, does not necessarily encompass every property motivating the research. This distinction is crucial as it underscores the meaningfulness of improvements and highlights the gap between what can be measured and what might be theoretically important. For instance, a judge's agreement on quality ratings can be quantified without capturing the nuanced interpretation of that quality, which may vary significantly between judges. Thus, focusing on a single operational quantity allows for clearer evidence and claims, even if broader constructs remain unresolved. A future study could retain these supporting dimensions alongside the primary summary, examining cases where they diverge. This approach avoids prematurely altering the primary interpretation based solely on whether one summary or another favors a preferred conclusion. By doing so, we ensure that a strong result on a bounded quantity remains valuable, even if the broader construct is not fully resolved. The scientific benefit lies in establishing a clearer correspondence between evidence and claim, rather than insisting on measuring every aspect of the subject.

Consensus should not manufacture a clean target

In the context of Judge Consistency and Judge Quality being distinct constructs, treating consensus as unqualified truth can lead to a model that appears independently validated, potentially ignoring the nuances behind the final decision. By preserving initial ratings, reasons for revision, and unresolved alternatives, future research can evaluate agreement with the final panel while also examining whether the procedure recognizes ambiguous cases. This approach ensures that the decision made by the panel does not erase the uncertainty relevant to understanding judges or automated assessments. It maintains a clean target for scientific claims about judges, avoiding the pitfalls of conflating reliability with validity.

Rubric structure and measurement

The intraclass-correlation (ICC) work highlights that different rating designs necessitate distinct reliability questions, as noted by Shrout and Fleiss (1979). For instance, a single-assessor score and a panel average are fundamentally different measures. When evaluating judges' consistency using an ICC, it's crucial to specify the design and intended use of the rubric structure, as outlined in the ICC reporting guideline by Koo and Li (2016). Failing to do so can lead to ambiguity. Readers cannot determine which reliability question was addressed.

Cohen's agreement coefficient (Cohen, 1960) is particularly useful for nominal classifications but should not be applied without careful consideration. It’s important to retain the category definition and prevalence context when interpreting agreement summaries, as Cohen’s formula assumes a chance-agreement baseline that may not apply to continuous scores or different rating designs. Thus, the review must critically assess whether the rubric structure aligns with the intended use of the data and maintain clarity in its reporting.

The decision pathway deserves its own evaluation

The decision pathway deserves its own evaluation, as it can offer insights into the effectiveness of judgments beyond just their outcomes. Observations alone might suggest a useful estimate but fail to address the appropriateness of subsequent actions. For instance, judges might agree on a quality score yet interpret this in ways that reflect personal biases rather than objective standards. A thorough assessment should consider not only the initial interpretation but also the decision-making process following it. This includes evaluating whether further review or deferral is necessary based on additional information. Simply repeating the analysis without considering these steps can miss critical nuances, such as how different judges might interpret similar evidence differently due to varying personal biases or interpretations. By examining the entire pathway from observation through interpretation and response, we can better distinguish between a useful signal in the data and an effective decision-making process. This approach maintains focus on the original scientific question while providing a more comprehensive understanding of judgment processes.

Criterion weights express a contest objective

A composite score can reflect strengths compensating for weaknesses, with weights indicating which criteria are most important for the contest's objective. When a ranking reverses under a different defensible weighting rule, it suggests a substantive tradeoff rather than a computational error. This distinction is crucial as it separates an uncertain judgment of one criterion from disagreement about the overall objective. Internal-consistency measures like Cronbach’s alpha do not resolve whether the assessment identifies a consistently reliable judge or one whose apparent quality hinges on specific scoring conventions. Thus, the measurement interpretation must make this tradeoff explicit, distinguishing between objective contest criteria and subjective judgments that may vary based on scoring preferences.

Statistical interpretation of assessor evidence

Coefficient alpha is indeed a measure of internal consistency, as Cronbach (1951) describes, but it does not equate to agreement among judges or assessors. A rubric might have criteria that are related yet subject to different interpretations by raters, indicating that the internal structure of items and the reliability of judgments are distinct concerns.

Moreover, the categorical-agreement analysis by Landis and Koch (1977) underscores the need for separate evidence when assessing observer agreement. Agreement summaries can vary significantly based on the classification problem and its underlying population, highlighting the importance of distinguishing between disagreements over challenging cases and inconsistencies in applying a settled rule. This qualitative distinction is crucial for maintaining clarity in evaluating judge consistency versus judge quality within a contest judgment context.

Reliability concerns a defined rating design

The intraclass-correlation (ICC) perspective, as articulated by Shrout and Fleiss (1979), and reporting guidelines provided by Koo and Li (2016), emphasize that the rating design is central to understanding reliability. When an ICC coefficient is reported without specifying which judge or panel was involved, it leaves readers uncertain about what aspect of agreement was measured. For instance, if a single judge's decisions are being assessed, this does not necessarily reflect the consistency between different judges or a panel average. Similarly, the reliability among observed judges does not inherently describe how a new judge would perform under that same rating design. Therefore, it is crucial to clarify these distinctions when reporting reliability measures. Doing so transforms a simple statement about judge quality into an interpretable claim about repeatability under a specified design rather than a general assessment of consistency across all possible scenarios. This approach ensures that the reliability coefficient accurately represents the intended measurement context.

Severity and inconsistency are not interchangeable

Severity and inconsistency are not interchangeable constructs, as demonstrated by judges who can score consistently higher than their peers while maintaining similar entry rankings. This phenomenon is exemplified by a judge who scores systematically higher but still agrees on the overall quality of entries, whereas another judge might have an average score but exhibit inconsistent judgments on specific cases. The conceptual distinction between these patterns is crucial. Systematic differences in scoring do not equate to consistency in judgment quality.

Absolute score thresholds can be misleading indicators of severity, as they may not account for the variability in how judges interpret and assess individual entries. In contrast, an ordering decision based on these scores might be less sensitive to the inherent variability in how judges rank entries. By retaining both systematic tendencies and residual disagreement, assessments can better reflect the complexity of judging tasks. This approach also necessitates a careful examination of whether apparent severity is due to the difficulty of the entries assigned rather than a flaw in the assessment process itself. Such feedback should serve as a diagnostic tool for identifying measurement issues, rather than an unexplained ranking of assessors.

A quality score has a development history

When assessing a quality score's development history, it is crucial to recognize that its utility does not equate to its validity as a broader measure of judge quality. An assessor-quality model trained on ratings from a panel can be validated by comparing its judgments with the same panel, but this approach risks reinforcing conformity to the instrument used during development. Thus, the evaluation question should not be conflated with validation of a more comprehensive construct. Instead, it is essential to maintain separation between feedback used to refine the model and observations reserved for assessment. This separation preserves the integrity of the scoring system by avoiding the risk of rewarding adherence to an instrument that may not fully capture broader judge quality. Additionally, examining sensitivity to defensible criterion weights and managing multiple comparisons can provide a more nuanced understanding of the score's utility within its defined rating design, thereby preserving its reliability concerns.

Related criteria do not establish rater agreement

In the internal-structure perspective articulated by Cronbach (1951) and elaborated upon by Tavakol and Dennick (2011), reliability focuses on the relationships between items within a test or rubric. This is fundamentally different from assessing whether multiple judges consistently assign similar scores to the same entries. A rubric can contain closely related criteria, yet judges may apply these differently. By conflating these two types of evidence under the umbrella of reliability, we obscure the actual evaluation that was conducted.

High internal consistency might be indicative of redundant criteria rather than a comprehensive coverage of the intended quality construct. Conversely, low agreement could reveal ambiguous wording rather than an inherently poor judge. These distinctions are crucial for a more diagnostic evaluation and help reduce reliance on a single attractive summary statistic. Thus, treating reliability and rater agreement as distinct constructs is essential for a thorough understanding of judges' performance and consistency.

What the observation represents

The observation suggests that within a hypothetical chart, one entry represents the consistency among judges, while another entry signifies their quality judgment. In the first level of aggregation, we see how judges are consistent in their assessments, agreeing on certain aspects of cases despite potential biases. This consistency can be measured through reliability indices, such as inter-rater agreement statistics, which focus on the degree to which different raters apply the same criteria.

At a second level of aggregation, the chart reflects judgments about the quality of decisions made by judges. Here, the interpretation shifts towards assessing whether judges are consistent in their interpretations of legal standards and principles, or if they exhibit variability in how they assess the fairness and correctness of rulings. This distinction is crucial because it highlights that quality judgments can be consistent yet biased, whereas reliability concerns the uniformity of assessments under a fixed set of criteria.

The observation clarifies that consistency among judges, as measured by reliability, does not equate to the quality of their judgments, which may vary due to personal biases or differing interpretations. This distinction is vital for understanding the nuances in judicial evaluation and the challenges in achieving both consistency and unbiased quality assessments.

Conclusion

High internal consistency among judges might indicate redundant criteria rather than comprehensive coverage of quality, while low agreement could reveal ambiguous wording rather than inherent poor quality. These distinctions are crucial for a more diagnostic evaluation, reducing reliance on a single summary statistic. Thus, treating reliability and judge quality as distinct constructs is essential for a thorough understanding of judges' performance and consistency. The consequence is that consistency among judges does not necessarily imply fair or accurate quality judgments, highlighting the need for separate metrics to assess both. An open question is how to better integrate these concepts, perhaps through a hybrid approach that combines reliability indices with quality assessments, to provide a more comprehensive evaluation of judicial performance.

References

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

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

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

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

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

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

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

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.

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

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