Studies in Knowledge and Innovation
Vol. 2 No. 10A (2026)

Internal Consistency Does Not Establish Agreement between Judges

Mark Wells[1] · Harry Gray[2] · Ryan King[3]

[1] Barbados · [2] Sierra Leone · [3] Lebanon

Published: October 26, 2026 | Article 1028

Abstract

Internal consistency describes relationships among items or criteria. It does not directly establish that different judges assign equivalent scores to the same entry.

Internal Consistency Does Not Establish Agreement between Judges

Abstract

Internal consistency describes relationships among items or criteria. It does not directly establish that different judges assign equivalent scores to the same entry.

Introduction

The evaluation system associated with Zhao et al. (2022) may contain several criteria whose scores move together. That pattern can reflect a common construct, overlapping criteria, or a general impression influencing several ratings. A review should avoid treating it as a complete test of assessor agreement. The latter requires a comparison of judgments across assessors for a defined target. High internal consistency can coexist with systematic differences between judges, while low consistency can occur when a rubric intentionally covers distinct dimensions. Naming the relationship being evaluated keeps a familiar psychometric summary from answering a different judging question by implication.

Assignment and structured variation precede adjustment

The mixed-model perspectives in Laird and Ware (1982) and Bates et al. (2015) suggest that when judges are assigned to different populations of entries, their unadjusted average scores can reflect both severity and case difficulty. This structured variation means that the grouping structure should align with the assignment design. If judges see distinct populations, their unadjusted scores might not accurately represent agreement between them. A fitted adjustment is only as interpretable as the design and assumptions supporting it. To avoid conflating data volume with the number of independent challenges to quality interpretation, it is crucial to report uncertainty from the relevant independent units rather than treating each rating as a separate assessor. This distinction helps maintain the integrity of the assessment and ensures that agreement between judges is not misattributed to differences in severity or case difficulty.

Rating designs and agreement

The distinction between explaining and predicting, as outlined by Shmueli (2010), underscores why a fitted model alone does not necessarily reveal the underlying mechanism or causal relationship (Shmueli, 2010). While a review might emphasize predictive success, it is crucial to acknowledge the uncertainty surrounding the substantive explanation of observed relationships. This context aligns with the notion that internal consistency, which measures the reliability among items or criteria, does not directly establish agreement between judges (Manuscript title: Internal Consistency Does Not Establish Agreement between Judges).

Turning to Laird and Ware's (1982) discussion on random-effects frameworks for longitudinal data, they emphasize the importance of interpreting grouping terms in relation to the process that created these groups. This is relevant when considering rating designs, where different judges may be assigned to varying conditions or categories. In such cases, agreement between judges should not be assumed based solely on internal consistency but must align with the specific design and purpose of the ratings.

Bates et al.'s (2015) lme4 paper highlights that successful estimation of mixed-effects models does not automatically validate the chosen model structure against scientific questions. The structure of the model should be guided by the allocation of subjects, the dependence among observations, and the intended

generalization beyond the data at hand. This approach ensures that the rating design is appropriately aligned with the research goals rather than being dictated by software convenience.

Agreement does not define a good rubric

Judges can agree on a rubric due to shared understanding and clear criteria, yet this agreement doesn't necessarily mean that the judges are assigning equivalent scores. The rubric might allow for multiple interpretations that lead to similar outcomes without being fundamentally aligned (Shmueli, 2010). Agreement does not establish that the rubric itself is a good measure of the intended quality or characteristics. Instead, reliability and construct validity should be evaluated separately. Reliability ensures consistency in scoring across judges, while construct validity examines whether the rubric measures what it claims to measure. Agreement between judges doesn't automatically imply that all interpretations are equivalent or that the rubric aligns with the intended criteria. If judges disagree due to different interpretations of the rubric, this disagreement can highlight potential ambiguities and tradeoffs in the assessment process. Evaluating judge agreement without considering the rubric's clarity and alignment with the substantive objectives of the assessment can lead to a misleading sense of consistency and validity. Therefore, developing a clear, unambiguous rubric is crucial for ensuring that judges are actually measuring what they intend to measure.

Rubric structure and measurement

Rubric structure and measurement is crucial for understanding how judges assign scores, especially when these scores are intended to be equivalent across different entries or assessments. The intraclass-correlation (ICC) literature, as discussed by Shrout and Fleiss (1979), underscores this point. While a single assessor's score represents one perspective, the ICC emphasizes that comparing average scores from multiple judges is a different measurement altogether. This distinction highlights that internal consistency measures the reliability of items or criteria within a single assessment, not necessarily the agreement between different judges scoring the same entry.

The ICC guideline (Koo & Li, 2016) further reinforces this by emphasizing the importance of specifying the model, the scale of measurement, and the definition of agreement in reliability statements. Simply reporting an ICC coefficient without context can obscure which aspect of reliability is being assessed. For instance, if a study uses a single-rater scoring method but claims internal consistency, it might not be providing a comprehensive measure of agreement between judges. Therefore, maintaining clarity about the rubric structure and measurement is essential for accurately interpreting and applying internal consistency measures in conceptual research reviews.

Selection is part of the evidence pathway

Selection of objects for reporting is a critical step that influences the interpretation of research findings. When an article selects one result from multiple possibilities, it introduces a distinction between the selected object and the process used to select it. This selection can be seen as part of the evidence pathway, where the chosen object (e.g., a specific judge's ratings) and the selection procedure are distinct entities.

When evaluating such a selection, researchers should distinguish between assessing whether the selected object itself is useful or how reliably the selection process produces such objects. For instance, if an article selects one set of ratings from various judges' assessments, it can be questioned whether these particular ratings are representative or if the selection process itself is flawed.

Confusion often arises when evidence that supports the validity of the selected object (e.g., internal consistency) is used to argue for the reliability of the entire selection procedure. By recording the decision pathway, researchers ensure that unsuccessful and unresolved cases are not dismissed but instead provide valuable insights into the selection process's behavior. These cases highlight potential flaws in the selection method rather than being ignored as anomalies.

Criterion weights express a contest objective

Criterion weights indeed express the contest's objective, as they prioritize certain criteria over others according to the goal at hand. When different judges apply these weights to the same entry, internal consistency measures may show that they are assigning equivalent scores, but this does not necessarily confirm agreement on the contest's objective. Internal-consistency evidence, such as Cronbach's alpha, indicates how consistently judges score items or criteria, but it does not address whether these scores reflect a fair and equitable evaluation of the entry. For instance, if one judge prioritizes creativity over technical proficiency while another places greater emphasis on technical accuracy, internal consistency might suggest that both judges are assigning equivalent scores to the same entry, yet they may fundamentally disagree on what constitutes a good submission. Thus, it is crucial to distinguish between an uncertain judgment of one criterion and disagreement about the contest's objective itself. Internal-consistency evidence does not settle the question of whether the assessment identifies a broadly reliable assessor or one whose apparent quality depends on a particular scoring convention.

Statistical interpretation of assessor evidence

Cohen's kappa coefficient, while useful for nominal classifications, does not directly assess agreement among judges when applied to continuous scores or different rating designs (Cohen, 1960). This statistical tool is prone to misinterpretation without careful consideration of the context and category definitions. On the other hand, Cronbach's alpha measures the internal consistency of a test's items but fails to establish agreement between assessors (Cronbach, 1951). Agreement among judges is a separate construct that requires its own evidence, such as explicit rubric criteria and structured variation precede adjustment procedures, ensuring that assessors apply these criteria consistently.

Moreover, Landis and Koch’s categorical-agreement analysis offers another perspective on interpreting observer agreement (Landis & Koch, 1977). Agreement summaries depend on the specific classification problem and its population context. Therefore, it is crucial to preserve this distinction between disagreement about difficult cases and inconsistency in applying a settled rule. In conclusion, while internal consistency is an important measure of test reliability, it does not establish agreement among judges, emphasizing the need for separate evidence and careful interpretation.

Consensus should not manufacture a clean target

The notion of consensus, while valuable for resolving immediate disputes, can inadvertently oversimplify the complexities inherent in measurement processes. By treating final decisions as unqualified truths, models that mimic these decisions can falsely appear independently validated. This is particularly problematic when using categorical-agreement frameworks like those proposed by Landis and Koch (1977), which emphasize the importance of understanding the nuances behind summary judgments. When a panel resolves disagreements, it does so within the context of specific criteria and definitions, leaving unresolved cases that are crucial for scientific claims about judges or automated assessment systems. A more rigorous approach would involve preserving initial ratings, reasons for revision, and unresolved alternatives. This allows for a deeper examination of agreement with the final decision while also acknowledging the uncertainty in ambiguous cases. By doing so, we prevent final decisions from erasing the very uncertainty that is essential for understanding the reliability and validity of assessment processes.

Severity and inconsistency are not interchangeable

Severity and inconsistency are not interchangeable, as demonstrated by differing patterns of systematic differences among judges. In one scenario, a judge might consistently rate entries higher than peers while maintaining the same relative orderings. Alternatively, a judge could have an average score similar to others but exhibit inconsistent scoring on specific entries. These distinctions are crucial because internal consistency measures do not directly assess whether different judges assign equivalent scores to the same entry. For instance, Shrout and Fleiss (1979) and Koo and Li (2016) emphasize that while systematic differences are important, they should not overshadow residual disagreement in rankings. Thus, a future assessment should consider both systematic tendencies and residual disagreements, acknowledging the inherent uncertainty of the assignment design. Additionally, it is essential to scrutinize whether apparent severity in scoring patterns reflects the difficulty of the entries assigned, rather than an unexplained ranking of assessors. By doing so, feedback to judges can be reframed as a diagnostic tool for identifying specific measurement issues rather than merely an evaluation of relative performance.

Reliability concerns a defined rating design

The reliability of judges is fundamentally linked to the specific rating design they employ, as highlighted by Shrout and Fleiss (1979) and Koo and Li (2016). When an intraclass-correlation coefficient is reported without specifying whether it refers to a single judge or a panel, it becomes ambiguous. A single judge's performance does not equate to the average of multiple judges, nor does the reliability of one set of judges necessarily translate to another set under different conditions. Thus, when discussing reliability, it is crucial to clarify whether the coefficient reflects a fixed number of judges or an average over a panel, and under what specific design these judgments were made. This clarity transforms reliability from a vague descriptor of judge quality into a precise measure of consistency within a defined rating system.

A quality score has a development history

In an assessor-quality model trained on panel ratings, agreement with the same panel can be misleading, as it does not establish independent validation of a broader construct. The development history of the quality score, which involves feedback used to revise the model and observations reserved for assessment, should be clearly distinguished. This separation ensures that assessments are not conflated with revision processes, preserving the integrity of both. Furthermore, examining sensitivity to defensible criterion weights and managing multiple comparisons can provide a more nuanced understanding of assessors' performance, aligning with recommendations from RUBIN (1976) and Benjamini and Hochberg (1995).

What the observation represents

In the hypothetical chart, one entry represents a specific instance of assessment or evaluation, encapsulating individual observations or judgments from a single judge. At the second level of aggregation, this unit is expanded to encompass multiple entries, reflecting the collective judgments of several assessors. This aggregation changes the interpretation because it shifts the focus from singular assessments to broader patterns of agreement or disagreement among judges. In the first level, we observe the variability and consistency within a single set of ratings. At the second level, we examine how these internal consistencies across different judges contribute to overall agreement or inconsistency in the rating system. Hence, understanding both levels is crucial for accurately interpreting the reliability and validity of assessment systems.

Conclusion

One consequence of this distinction is that it highlights the need for careful interpretation of internal consistency measures. For instance, judges may have different levels of severity in their scoring without necessarily assigning equivalent scores to the same entry. This means that internal consistency does not directly establish agreement between judges. An open question is whether the observed differences in severity among judges are due to the difficulty of the entries themselves or to systematic differences in their scoring patterns. Further research should consider both systematic tendencies and residual disagreements to better understand the reliability of judges.

References

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

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

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

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

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

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

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.

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

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

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