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

Fair Comparison of Judges across Different Entry Assignments

Ethan Carter[1] · Owen Stone[2] · Ryan King[3]

[1] Kyrgyzstan · [2] Anguilla · [3] Japan

Published: November 02, 2026 | Article 1033

Abstract

Judges assigned different work do not necessarily face comparable evaluation tasks. Fair assessor comparisons need an account of entry difficulty and assignment overlap.

Fair Comparison of Judges across Different Entry Assignments

Abstract

Judges assigned different work do not necessarily face comparable evaluation tasks. Fair assessor comparisons need an account of entry difficulty and assignment overlap.

Introduction

A raw summary of judge performance in the setting of Zhao et al. (2022) can mix assessor behavior with differences in assigned entries. A review should identify what shared basis makes two summaries comparable. Adjustment may help express a target comparison, but cannot create information absent from a disconnected assignment structure. The distinction between an observed panel and a hypothetical reassignment should remain explicit. A fair interpretation could consequently be narrower than a complete ranking of assessors. It may establish which comparisons are supported and leave others unresolved, rather than presenting every calculated score as equally informative about judge quality.

Severity and inconsistency are not interchangeable

Severity and inconsistency are not interchangeable terms, even though they may seem similar. A judge might exhibit a consistent tendency to score higher than peers while maintaining the same relative order of entries, or conversely, score an average but display significant disagreement on specific cases. These patterns can be subtle and are often masked by a single quality assessment.

Absolute scores, which indicate a fixed numerical value, can vary in sensitivity depending on the difficulty of the entries assigned. A higher score does not necessarily mean better performance. It could simply reflect the inherent challenges of certain tasks. Similarly, an average score may mask significant variations in judgment quality across different cases. Residual disagreement among judges, or the variability beyond a consistent tendency, is another important aspect to consider.

To address these complexities, future assessments should account for both systematic tendencies and residual disagreements, acknowledging the inherent variability of judging tasks. This approach would provide a more nuanced understanding of judge performance and help diagnose specific measurement issues rather than merely comparing rankings of assessors.

Rating designs and agreement

In the context of rating designs and agreement, differing assignments for judges can lead to inconsistencies in evaluations, undermining fair comparison. To mitigate this, it is crucial to account for entry difficulty and assignment overlap. Without these considerations, rating scales and criteria applied by different judges will not be comparable, diminishing the validity of assessment outcomes. For example, if one judge evaluates entries that are predominantly simple and straightforward, while another faces a mix of complex and challenging tasks, their ratings will likely diverge significantly. To ensure fair comparison, standardizing tasks and criteria across all entries and assigning them to judges in a way that minimizes overlap and maximizes variety is necessary. This requires careful design of entry assignments and thorough consideration of how these assignments influence the evaluation process.

Agreement does not define a good rubric

In the context of fair comparison of judges across different entry assignments, agreement among judges can be indicative of a shared, clear rule but may overlook critical aspects critical to the contest. Conversely, disagreement among judges often arises from multiple plausible interpretations of the rubric, highlighting distinct concerns of reliability and construct validity. While agreement ensures consistency, it does not guarantee a robust or fair evaluation system. A good rubric should address the inherent differences in entry difficulty and assignment overlap, as suggested by Shmueli (2010). Agreement alone is insufficient to establish construct validity, emphasizing the need for a rubric that accounts for these nuances. Developing such a rubric avoids conflating severity and inconsistency, thereby providing a fairer comparison of judges across various assignments. Ultimately, a rigorous evaluation requires a rubric that is not only clear but also sensitive to the complexities of different entry types and assignment contexts.

Rubric structure and measurement

Rubric structure and measurement is crucial for fair comparison among judges handling different entry assignments. When judges assess tasks of varying difficulty, it's essential to design a rubric that accounts for these differences. The text references partial pooling, which stabilizes estimates while adjusting how local variation is represented. This implies that a well-structured rubric should allow for adjustments in scoring based on the complexity of the entries. Additionally, the missing-data framework emphasizes the importance of understanding the observation process. Inconsistent or missing data can lead to biased assessments if not handled appropriately. Therefore, a comprehensive rubric design must consider how observed values relate to the complete population, ensuring that any conclusions drawn are valid and fair across different entry assignments. By addressing these issues, judges can provide more comparable evaluations, facilitating a fairer comparison of their performance.

A useful quantity need not be a complete construct

An operational quantity, such as a specific scoring metric, is selected because it can be observed and analyzed. The usefulness of this quantity does not imply that it captures every property that motivated the research, which underscores the importance of understanding its limitations. For example, when comparing judges on different entry assignments, an improvement in one aspect may not reflect improvements across all dimensions. This makes the meaning of an improvement clearer but does not demand that every study measure every aspect of its subject.

In future assessments, maintaining supporting dimensions alongside a primary summary can help identify cases where these dimensions disagree. Avoiding changes to the primary interpretation only after seeing which summary favors a preferred conclusion ensures that the core findings remain robust and interpretable. A strong result on a bounded quantity, such as a particular scoring method, can still be valuable even if the broader construct is unresolved. The scientific gain here is a clearer correspondence between evidence and claim, not a demand that every study measure every aspect of its subject.

Criterion weights express a contest objective

Criterion weights are crucial in defining the contest's objectives, as they indicate which strengths and weaknesses can be substituted for each other. When judges evaluate entries of varying difficulty, these weights ensure that comparisons between them are fair. A composite score, with its assigned weights, allows certain strengths to compensate for weaknesses, but it also reveals substantive tradeoffs that might not be apparent otherwise. If a ranking reverses under another defensible weighting rule, this suggests that the contest's objective is being compromised rather than the computation itself. The measurement interpretation must make these tradeoffs explicit, distinguishing between an uncertain judgment of one criterion and disagreement about the objective itself. Internal-consistency evidence, such as Cronbach's alpha, does not resolve this debate. It can only indicate reliability within a consistent scoring convention. Thus, clarity on the contest's purpose is essential to avoid conflating quality judgments with the conventions used in evaluation.

Statistical interpretation of assessor evidence

The ICC reporting guideline emphasizes the importance of specifying rating units, model, and agreement definitions for reliability statements (Koo & Li, 2016). Failing to do so leaves readers uncertain about which aspect of reliability is being reported. Coefficient alpha, a measure of internal consistency in test items (Cronbach, 1951), does not equate with rater agreement. A rubric can have criteria that are related yet be applied differently by assessors, meaning item consistency and rater reliability should be evaluated separately.

Similarly, categorical-agreement analysis offers another framework for interpreting observer agreement (Landis & Koch, 1977). Agreement summaries depend on the specific classification problem and its population. Thus, a synthesis should preserve the distinction between disagreement over difficult cases and inconsistency in applying a settled rule. In both contexts, it is crucial to recognize that these distinctions are separate from overall test reliability and item consistency.

Different scales need different agreement questions

The nominal-scale perspective, as highlighted by Cohen (1960), and the categorical-agreement perspective, as outlined by Landis and Koch (1977), focus on classification problems. These perspectives are applicable when judgments are binary or categorical, where the scale of disagreement is straightforward. However, for continuous ratings or ordered criteria, a different approach to measuring agreement is necessary. Herein lies the issue: different scales necessitate different questions about agreement.

The population of entries also plays a critical role in assessing judges' performance. If the set of entries is narrow and varied, judges may appear more consistent due to large differences among individual cases. Conversely, if the set is broad and similar, judges might seem less consistent because there are fewer differences to observe. Neither approach is universally correct. Thus, specifying the intended case distribution is essential. An aggregate coefficient should be interpreted as a summary of that specific distribution rather than a universal measure of judge performance independent of the entries. This distinction ensures that comparisons between judges are fair and meaningful, reflecting the variability and characteristics of the entries they evaluated.

Reliability concerns a defined rating design

The reliability of judges, as defined by the intraclass-correlation coefficient (ICC) and as reported in Shrout and Fleiss (1979) and Koo and Li (2016), hinges on the specific rating design used. This means that a single judge's performance or an average across a panel does not inherently reflect the reliability of a new judge population. The ICC can vary significantly depending on whether it is calculated for a single judge, a panel, or even if systematic differences are accounted for. Without specifying these parameters, a reader cannot accurately interpret the reliability coefficient. For instance, if judges are compared based on their consistency in evaluating entries with varying levels of difficulty, this does not equate to comparing them against a panel where all entries are of uniform difficulty. Therefore, it is crucial to clarify the rating design when reporting reliability coefficients. This clarification would transform a concise measure of judge quality into an interpretable statement about repeatability under a defined design, thereby addressing concerns about comparability across different entry assignments.

A quality score has a development history

The development history of a quality score can be problematic if it is evaluated solely on agreement with the same panel that developed it. This raises concerns about whether the score truly represents independent validation of a broader construct, or merely a reinforcement of the evaluation instrument itself. To address this issue, a future workflow should separate feedback used to revise the model from observations reserved for assessment. This separation would preserve missing ratings and exclusions, allowing for a more nuanced understanding of how the quality score performs under different conditions. Additionally, examining sensitivity to defensible criterion weights and managing multiple comparisons through methods like Benjamini and Hochberg (1995) can provide a more robust framework for assessing assessors. By doing so, the evaluation question can be reframed to avoid the pitfalls of conflating conformity with independent validation.

Assignment and structured variation precede adjustment

When assessing judges across different entry assignments, it is crucial to account for the varying difficulty levels and overlap between tasks assigned to each judge. Laird and Ware (1982) and Bates et al. (2015) emphasize that mixed-model perspectives are essential when entries, judges, and occasions contribute distinct variation. If judges are evaluating different populations of entries, their unadjusted average scores can inadvertently mix severity with case difficulty. This is because the grouping structure should follow the assignment design, ensuring that each judge’s performance is appropriately compared to the tasks they were given.

To avoid misinterpretation, it is important to report uncertainty from the relevant independent units rather than treating every rating as a separate assessor. This approach keeps data volume distinct from the number of independent challenges to the proposed quality interpretation. By focusing on the assignment and structured variation that precedes adjustment, we can provide a fairer comparison of judges, ensuring that their performance is accurately assessed based on the tasks they were assigned.

What the observation represents

The observation represented by one entry in a hypothetical chart of judges' evaluations suggests a unit that can vary significantly based on the specific task assigned. In the first level of aggregation, we consider the individual entries-judges' ratings for a particular assignment. These ratings reflect different challenges and complexities inherent to each entry, such as varying difficulty levels or unique aspects of the work being evaluated.

At a second level of aggregation, however, we interpret these individual entries collectively, noting that they are part of a larger, more comprehensive evaluation task. This broader perspective allows us to account for assignment overlap and difficulty, recognizing that judges may face tasks with similar objectives but differing degrees of complexity or scope. By aggregating these entries, we can better understand the overall consistency and fairness across different assignments, ensuring that comparisons between judges are not skewed by the varying nature of their work.

Conclusion

One consequence of not accounting for the varying difficulty levels and overlap between tasks assigned to judges is misinterpretation of their performance. For example, if judges are evaluating different populations of entries, their unadjusted average scores could inadvertently mix severity with case difficulty. An open question in this field is how to best incorporate the assignment and structured variation into a fairer comparison of judges, ensuring their performance is accurately assessed based on the tasks they were assigned.

References

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

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

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

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

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

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

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.

Gelman, A., Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. . https://doi.org/10.1017/cbo9780511790942