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

A Research Agenda for Auditable Data-Driven Contest Assessment

Austin Chapman[1] · Brian Elliott[2] · Jason Reed[3]

[1] Falkland Islands (Malvinas) · [2] Saint Kitts and Nevis · [3] Solomon Islands

Published: November 05, 2026 | Article 1038

Abstract

An auditable judging system makes the path from rubric to decision inspectable. The most useful next evidence addresses an uncertainty in that path rather than adding another unexplained summary score.

A Research Agenda for Auditable Data-Driven Contest Assessment

Abstract

An auditable judging system makes the path from rubric to decision inspectable. The most useful next evidence addresses an uncertainty in that path rather than adding another unexplained summary score.

Introduction

The data-driven topic of Zhao et al. (2022) can be organized around the links between criteria, assignments, individual ratings, aggregation, and the final contest decision. A review should identify where a dispute about quality enters that chain. Additional observations are valuable when they distinguish the competing interpretations at that point. A new aggregate statistic may improve description without resolving the dispute. The agenda is therefore about evidential purpose as much as data volume. Making the decision pathway explicit permits both useful conclusions and clear statements of what the available rating record still cannot establish.

Related criteria do not establish rater agreement

Cronbach (1951) and Tavakol and Dennick (2011) discuss internal consistency, which assesses the relationships among items within a rubric. This differs from assessing whether multiple judges give similar scores to the same entry. A rubric can have criteria that are closely related but still be applied differently by different raters. By conflating these two types of evidence under one reliability label, we risk obscuring what was actually evaluated.

High internal consistency might indicate that the criteria are redundant rather than providing a comprehensive coverage of the intended quality construct. Conversely, low agreement between judges could highlight ambiguous wording or unclear criteria rather than an inherently poor judgment. These distinctions provide a more diagnostic evaluation and reduce our reliance on a single attractive summary statistic.

Rating designs and agreement

In the context of rating designs and agreement, the distinction between explaining and predicting is crucial for understanding the path from rubric to decision. While a fitted model may indicate predictive success, it does not necessarily clarify the mechanism or causal explanation behind observed relationships (Shmueli, 2010). This underscores the need for a focused examination of how ratings are assigned and agreed upon, rather than merely focusing on the accuracy of predictions.

The random-effects framework for longitudinal data offers insight into structured variation within groups (Laird & Ware, 1982). The interpretation of grouping terms hinges on the process that created these groups. In contest assessment, this means considering how ratings are assigned and how they align with the intended criteria or rubric. Without a clear understanding of the rating design and the agreement among raters, it is difficult to discern whether observed patterns in scores are due to the intended criteria or other factors.

The computational framework provided by lme4 (Bates et al., 2015) is valuable for fitting linear mixed-effects models. However, successful estimation alone does not provide evidence that the chosen model structure answers the scientific question of rating agreement. The model structure should be guided by the allocation of ratings, the dependence among raters, and the intended generalization of findings to

broader contexts. This ensures that the models are not merely convenient but also aligned with the research objectives and practical applications in contest assessment.

The decision pathway deserves its own evaluation

The argument that the decision pathway deserves its own evaluation is bolstered by the need for a thorough understanding of the transition from rubric to decision, which can be obscured by focusing solely on summary scores. An audit of this pathway allows us to discern whether actions are justified based on the evidence provided or if further scrutiny is necessary. By examining the entire process-ranging from observations to interpretations and finally decisions-researchers can identify inefficiencies or biases that might not be apparent when relying on a final, aggregated score alone. For instance, it could reveal instances where additional review or deferral is warranted, ensuring that decisions are not made hastily or without sufficient data. This comprehensive evaluation ensures that the decision-making process is both transparent and effective, aligning with the goal of making judging systems more auditable and fair.

Criterion weights express a contest objective

Criterion weights are crucial for specifying the contest's objectives, as they determine which strengths can compensate for weaknesses and vice versa. This means that if a ranking is reversed under another defensible weighting rule, it exposes a substantive tradeoff rather than an error in computation. The measurement interpretation must highlight these tradeoffs to make them visible. However, distinguishing between uncertain judgments of one criterion and disagreements about the contest's objective itself remains challenging. Internal-consistency evidence, such as Cronbach’s (1951) reliability index, does not settle whether the assessment identifies a broadly reliable judge or one whose apparent quality hinges on a particular scoring convention. This distinction is necessary to clarify the purpose of the assessment and its validity.

Rubric structure and measurement

The rubric structure and measurement are foundational to an auditable judging system, as they directly impact how decisions are made based on evaluative criteria. The use of partial pooling through regression and multilevel modeling (Gelman & Hill, 2006) is particularly relevant here, as it allows for the stabilization of estimates while adjusting for local variation within the rubric's structure. This approach ensures that the system remains robust and adaptable to varying conditions without losing the nuanced details inherent in subjective assessment criteria.

However, the missing-data framework (RUBIN, 1976) also plays a crucial role in understanding the limitations and assumptions of such models. Simply accepting incomplete data does not inherently make it harmless. Rather, it necessitates careful explanation of how observed values relate to the complete population being assessed. This synthesis should articulate the specific ways in which the observed data points are connected to the broader context, thereby enhancing transparency and accountability in the decision-making process.

Furthermore, adhering to ICC reporting guidelines (Koo & Li, 2016) is vital for ensuring clarity and precision in reliability statements. By specifying the model type, rating units, and agreement definitions, these guidelines help readers understand the specific questions being addressed and the populations to which uncertainty applies. This attention to detail is essential for building an auditable system where every aspect of data analysis and interpretation can be scrutinized and validated.

Consensus should not manufacture a clean target

The concept of consensus in judging systems can inadvertently reinforce an unqualified truth, masking the complexity and variability inherent in the decision-making process. By treating consensus as a definitive endpoint, models that mimic panel decisions may appear independently validated, obscuring the nuances and disagreements that are central to understanding the measurement process. A more nuanced approach would preserve the initial ratings and reasons for revision, allowing for a comprehensive evaluation of agreement with the final panel. This method acknowledges the complexity of ambiguous cases and ensures that the decision pathway is not eroded by a clean target. By preserving these details, future research can better assess how contest decisions are made and what they reveal about the judges' criteria and the measurement process itself.

Statistical interpretation of assessor evidence

Cohen's agreement coefficient, while useful for nominal classifications, is not suitable for continuous scores or alternative rating designs (Cohen, 1960). To properly interpret its summary, it requires a clear category definition and context regarding prevalence (Cohen, 1960). The statistical interpretation of assessor evidence should therefore focus on these foundational elements to avoid misapplication.

In contrast, Cronbach's alpha offers insights into the internal consistency of a test or scale but does not independently validate the construct or judgments made from it (Tavakol & Dennick, 2011). For an auditable judging system, the reliability and validity of the assessment process must be established through a comprehensive analysis that includes both categorical agreement and internal consistency measures.

Landis and Koch's work on observer agreement highlights the importance of distinguishing between disagreement about difficult cases and inconsistency in applying a settled rule (Landis & Koch, 1977). This distinction is crucial for understanding the reliability of assessor evidence. By preserving these distinctions, an auditable system can provide a more robust foundation for decision-making, ensuring that disagreements are not conflated with systematic errors in judgment.

Assignment and structured variation precede adjustment

An audit of judge scores should establish which entries were rated more than once and who rated them on each occasion. Laird and Ware (1982) and Bates et al. (2015) provide statistical frameworks for representing dependence among related observations. In a contest, the assignment record determines which of those relationships can inform an adjustment. If one panel sees only strong entries and another sees only weak entries, their mean scores reflect the material assigned to them as well as their scoring tendencies. An adjusted comparison should identify any shared entries that inform it and explain where it relies on assumptions about the two panels' assignments.

Moreover, it is essential to report uncertainty from the relevant independent units rather than treating every rating as a separate assessor. This approach ensures that data volume does not obscure the distinct challenges faced by the quality interpretation being proposed. By maintaining a clear distinction between data volume and the number of independent assessments, we can better interpret the adjustments made and their implications on contest evaluation.

Severity and inconsistency are not interchangeable

Severity and inconsistency are not interchangeable concepts, as demonstrated by judges who can score consistently higher than peers without altering the overall entry rankings. Conversely, another judge might exhibit a more even score distribution but display significant disagreement on specific entries. These distinctions matter because they highlight different aspects of judging performance: one reflects systematic bias in scoring, while the other indicates variability in decision-making.

Absolute score thresholds may be influenced by the severity of entries, potentially leading to biased assessments. However, ordering decisions are not solely determined by these thresholds but also by the consistency and variability in judgment. A more nuanced approach would involve retaining both systematic tendencies and residual disagreements, thereby accounting for the assignment design's inherent uncertainty. This would allow for a clearer diagnosis of measurement issues rather than treating apparent severity as an unexplained ranking of assessors.

Agreement does not define a good rubric

Judges can agree on a rubric when they share a clear rule but this agreement does not ensure the rubric is a good one. The rubric might omit properties that influenced the contest's design, and judges may disagree if the rubric allows for multiple interpretations. Reliability and construct validity are distinct concerns. Reliability assesses how consistent judgments are, while construct validity checks whether those judgments align with the intended criteria. Therefore, separate arguments are needed to demonstrate a rubric's effectiveness. Shmueli (2010) suggests that reliability and construct validity should be evaluated independently, as agreement on a rubric does not necessarily imply it is well-designed. A future assessment could focus on cases where disagreement illuminates the rubric's strengths and weaknesses, rather than eliminating differences to achieve a cleaner label. This approach ensures that the instrument itself is scrutinized, not just individual assessors.

Different scales need different agreement questions

Cohen's (1960) nominal-scale approach and Landis and Koch's (1977) categorical-agreement perspective are indeed relevant for classification problems, but they may not fully capture the nuances of continuous ratings or ordered criteria. These scales require a different methodological framework to assess agreement, as they involve judgments that are not necessarily binary or categorical. In such cases, the interpretation of disagreement should be tailored to the specific characteristics of the data being evaluated.

Moreover, the composition of the population of entries is crucial for understanding the level of consistency among judges. A narrow set of entries might obscure variability within a contest, while a broad set could highlight differences that are meaningful in the context of the competition. Neither approach is universally applicable. Therefore, it is important to explicitly define the intended case distribution and consider any exceptions that may arise. By doing so, an aggregate coefficient can more accurately represent the variability within the specific population being assessed rather than a general measure of judge consistency independent of the entries.

What the observation represents

The hypothetical chart represents individual entries as scores on a rubric, each corresponding to a specific criterion or aspect of performance. At a second level of aggregation, these entries aggregate into overall ratings or decisions made by judges based on the rubric. The interpretation at both levels differs fundamentally: at the first level, each score is an assessment of a distinct attribute, whereas at the second level, the aggregated scores are summaries that reflect the collective judgment on a broader set of criteria. Thus, understanding how these scores aggregate into final decisions requires unpacking the nuances of how individual criteria contribute to overall judgments. This distinction highlights the complexity in translating detailed rubric assessments into overarching decisions, which is crucial for ensuring transparency and accountability in contest assessment systems.

Sorting records into comparable groups

When grouping records into comparable groups, it is plausible that inclusion criteria could alter the appearance of an observed association. For instance, consider a scenario where a rubric for assessing contest entries is structured to prioritize creativity and execution separately. If only entries with high creativity scores are included in one group and only those with high execution scores in another, the observed association between creativity and execution might appear stronger than it actually is. This is because the groups are not representative of the entire population of entries. They are skewed towards specific qualities. As a result, any observed positive correlation between creativity and execution could be an artifact of the grouping process rather than an intrinsic feature of the entries themselves. Thus, it is crucial to ensure that records are grouped in a way that maintains representativeness and allows for a more accurate assessment of the relationship under investigation.

Conclusion

One consequence of this approach is that the interpretation of disagreement in continuous or ordered criteria is more nuanced and context-dependent. This means that the variability within a contest can be more accurately understood by considering the specific characteristics of the entries and the criteria being evaluated. An open question is how to balance the need for general measures of judge consistency with the importance of understanding variability specific to the population of entries. This requires careful consideration of the composition of the population and the criteria being assessed, as a broad set of entries might highlight differences that are meaningful in the context of the competition.

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

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

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

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.

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

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

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

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

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

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