Operational Alignment of Confidence-Based Flagging Methods in Automated Scoring

Corey Palermo, Troy Chen & Arianto Wibowo

Abstract

In hybrid scoring systems, confidence thresholds determine which responses receive human review. This study evaluates a relative (within-batch) thresholding method against an absolute benchmark across ten items. Results show near-perfect agreement and modest distributional differences, supporting the relative method's validity as a scalable, operationally viable approach for flagging low-confidence responses.

1 Introduction

In large-scale summative assessment programs, hybrid scoring systems that combine automated engines with human raters are commonly used to balance efficiency and accuracy. To preserve human scoring resources while maintaining scoring validity, these systems often rely on a measure of model confidence to identify which responses should be routed to human reviewers. For example, in Measurement Incorporated's hybrid automated-human scoring system, each student response is first evaluated by a scoring engine that assigns both a rubric-based score and a continuous confidence value. This confidence value reflects how well the response aligns with patterns learned from previously human-scored examples. When confidence is high, the model's score is accepted; when confidence is low, the response is routed to an expert human rater for review and final score assignment. The engine's use of floating-point scores rather than discrete categories acknowledges that writing quality exists on a continuum. As such, low-confidence predictions often arise when a response falls between score points or exhibits features that are atypical relative to the model's training data.

In practice, we use a relative (within-batch) thresholding approach to determine which responses are flagged for human scoring. Because operational scoring occurs continuously over several weeks, responses are processed in discrete batches as tests are completed. The model evaluates scoring certainty within each batch and flags approximately 10% of responses reflecting the lowest confidence. This strategy enables consistent workload distribution for human raters, supports timely data delivery, and ensures manageable flow across the scoring window. By contrast, an absolute thresholding approach—which would require evaluating confidence relative to the full population of responses—poses logistical challenges, in particular delayed score reporting.

Although the relative approach offers clear operational advantages, it is not known how well it approximates the theoretical ideal of an absolute confidence threshold. The present study investigates the extent to which the relative (within-batch) thresholding approach provides a robust and valid method for identifying low-confidence responses.

The study addresses three research questions:

RQ1: To what extent do the relative and absolute methods identify the same responses as low-confidence?

RQ2: Do the responses flagged by the relative method exhibit similarly low confidence values compared to those flagged by the absolute method, in terms of their overall distribution?

RQ3: Do the responses flagged by the relative method differ in median confidence values from those flagged by the absolute method?

Through a series of statistical comparisons aligned with each research question, we examine the extent to which the relative method replicates the behavior and outcomes of an absolute thresholding approach. These analyses evaluate the overlap in flagged responses, the similarity in their confidence value distributions, and potential

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 56–60
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

differences in central tendency, providing a multi-faceted assessment of the relative method's robustness.

2 Methods

To address the study's research questions, we conducted a series of statistical analyses across ten items, each designed to evaluate a distinct aspect of the alignment between the relative (within-batch) thresholding method and an absolute confidence threshold.

To evaluate the extent to which the relative and absolute methods identify the same responses as low-confidence (RQ1), we conducted McNemar's tests to assess whether the proportion of discordant classifications—responses flagged by one method but not the other—differed significantly. To further quantify the degree of agreement, we calculated F1 scores and Cohen's kappa for each item.

To assess whether the relative method captures responses with similarly low confidence values as those flagged by the absolute method (RQ2), we conducted Kolmogorov–Smirnov (K-S) tests. These non-parametric tests compared the full distributions of raw confidence values for responses flagged by each method, providing a measure of distributional similarity without assuming a specific shape or variance structure.

Finally, to examine whether the responses flagged by the relative method differ in central tendency from those flagged by the absolute method (RQ3), we performed Mann–Whitney U tests. These tests specifically assessed whether there were significant differences in the median confidence values between the two groups, offering a complementary perspective to the K-S analyses focused on overall distribution.

Together, these methods provide a multi-dimensional evaluation of the relative thresholding approach's robustness and its alignment with the conceptual goals of confidence-based response flagging.

3 Results

3.1 RQ1

McNemar's tests were used to assess whether the relative and absolute methods differ in how they classify responses as low-confidence. For each item, a 2×2 contingency table was constructed, and the test evaluated whether the proportion of discordant cases—responses flagged by one method but not the other—was significantly asymmetric. As shown in Table 1, none of the McNemar tests reached statistical significance (all p-values > .95), indicating no evidence of systematic disagreement between the methods.

To complement these significance tests, F1 scores and Cohen's kappa values were computed

Item ID	Grade	N	McNemar χ^2	p-value	F1 Score (%)	Kappa
X01	8	74050	0.003	0.956	97.8	0.976
X02	4	71140	0.001	1.000	91.0	0.901
X03	7	73928	0.003	0.953	98.1	0.979
X04	8	73806	0.002	0.967	96.0	0.956
X05	5	73422	0.003	0.957	97.6	0.974
X06	3	70868	0.001	0.974	93.2	0.925
X07	6	73764	0.002	0.966	96.3	0.959
X08	6	73900	0.003	0.960	97.4	0.971
X09	5	73322	0.001	0.974	93.2	0.924
X10	7	73902	0.001	0.972	94.5	0.939

Table 1: Agreement between relative and absolute methods across items.

to quantify the degree of agreement. F1 scores ranged from 91.02% to 98.08%, reflecting a high degree of precision and recall across items. Cohen's kappa values, which adjust for chance agreement, ranged from 0.901 to 0.979, consistently exceeding the commonly cited threshold ($\kappa > 0.90$) for near-perfect agreement (Landis & Koch, 1977). Item X02 showed the lowest observed agreement (F1 = 91.02%, κ = 0.901), while item X03 showed the highest (F1 = 98.08%, κ = 0.979). These findings indicate that despite using different thresholding strategies, the relative and absolute methods align closely in practice, identifying largely overlapping subsets of responses for human scoring.

3.2 RQ2

To examine whether the confidence value distributions of flagged responses differed between the relative and absolute methods, Kolmogorov–Smirnov (K-S) tests were conducted for each of the ten items. Table 2 displays the K-S test statistics, sample sizes, and p-values for each item.

Statistically significant differences in the distributions were observed for eight of the ten items ($p < .05$), with K-S statistics ranging from 0.0236 to 0.0898. For the remaining two items (X01 and X03), the tests were not statistically

significant, indicating no detectable difference in confidence distributions between the two methods for those items.

Although statistical significance was common, the magnitude of the observed differences—as indicated by the K-S statistics—was consistently small, with all values falling below 0.10. In the context of the Kolmogorov–Smirnov test, the KS statistic represents the maximum vertical distance between the empirical cumulative distribution functions of the two samples. A value below 0.10 suggests that the most extreme divergence between the relative and absolute confidence distributions is less than 10% at any point along the confidence continuum. These values are often interpreted as indicating a negligible to small effect size, implying that while the distributions are not identical, the differences are modest and unlikely to meaningfully alter the classification of responses as low-confidence.

Item ID	Grade	N Flagged	K-S Statistic	p-value
X01	8	7552	0.022	0.052
X02	4	6796	0.090	< .001
X03	7	7605	0.019	0.121
X04	8	7360	0.040	< .001
X05	5	7430	0.024	0.032
X06	3	7106	0.068	< .001
X07	6	7309	0.037	< .001
X08	6	7484	0.026	0.012
X09	5	7084	0.068	< .001
X10	7	7449	0.055	< .001

Table 2: Kolmogorov–Smirnov test results comparing confidence distributions

To illustrate these patterns, Figure 1 displays histograms of confidence values for three representative items—one with no significant difference (X03), one with moderate divergence (X10), and one with the largest observed difference (X02). As shown, the distributions overlap substantially, with only minor shifts in the region of greatest density. These visualizations reinforce the conclusion that the relative method tends to flag responses from the same general region of the confidence distribution as the absolute method. While some divergence is detectable, the observed differences are limited and unlikely to compromise scoring validity.

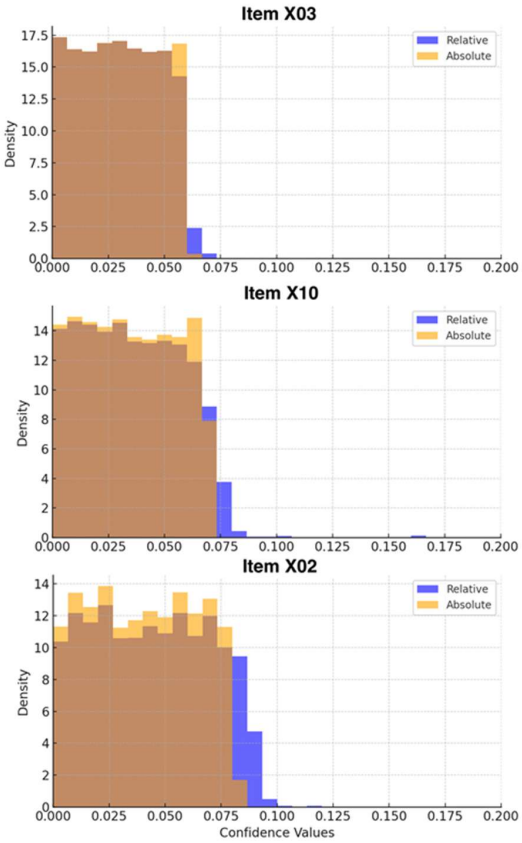

Figure 1: Histograms of confidence values: relative vs. absolute flagging.

3.3 RQ3

Table 3 presents the sample sizes, U statistics, *p*-values, and corresponding estimates of the Common Language Effect Size (Vargha & Delaney, 2000; McGraw & Wong, 1992), denoted as $\hat{P}_1$.

$\hat{P}_1$ represents the probability that a randomly selected response from one group (e.g., flagged by the absolute method) has a higher confidence value than a randomly selected response from the other group (e.g., flagged by the relative method). Under the null hypothesis of equal distributions, $\hat{P}_1 = 0.5$, indicating no systematic difference in central tendency. Values modestly above or below 0.5 suggest directional but generally small effects. Statistically significant differences in median confidence values were observed for six of the ten items, with p-values ranging from < .001 to .012. For the remaining four items (X01, X03, X05, and X08), results were not statistically significant, indicating no detectable difference in central tendency between the two groups of flagged responses.

The estimated $\hat{P}_1$ values across all items ranged from 0.501 to 0.543. These values suggest that even when differences were statistically

Item ID	Grade	N Flagged	U Statistic	p-value	$\hat{P}_1$
X01	8	7552	28700007	0.493	0.503
X02	4	6796	25070858	< .001	0.543
X03	7	7605	28975153	0.833	0.501
X04	8	7360	28008532	< .001	0.517
X05	5	7430	27915699	0.231	0.506
X06	3	7106	26815639	< .001	0.531
X07	6	7309	27639771	< .001	0.517
X08	6	7484	28466365	0.081	0.508
X09	5	7084	26357199	< .001	0.525
X10	7	7449	28402035	0.012	0.512

Table 2: Mann–Whitney U test results comparing median confidence values.

significant, the relative method only slightly increased the probability of flagging responses with higher confidence values compared to the absolute method. Importantly, all $\hat{P}_1$ values were greater than 0.5, indicating a consistent directional trend across items. This pattern aligns with the design of the relative method, which evaluates responses within batches; as a result, it may include some responses that exceed a fixed global threshold but still represent the lower-confidence tail within that batch.

Taken together, these findings reinforce the conclusion that the two methods target similar segments of the confidence distribution. While the relative method yields small, systematic shifts in central tendency compared to the absolute approach, these shifts are modest and consistent with its operational design. They do not undermine its ability to identify responses with genuinely low scoring confidence.

4 Discussion

This study evaluated the robustness of a relative (within-batch) thresholding method for identifying low-confidence responses by comparing it to an absolute thresholding approach across ten assessment items. The results indicate that although the two methods use different reference frames for determining confidence, they yield closely aligned outcomes in practice. McNemar's tests showed no statistically significant differences in flagging decisions across any item, indicating that the two methods do not systematically disagree in their classifications. Agreement metrics further reinforced this pattern, with F1 scores above 91% and Cohen's kappa values consistently exceeding 0.90—benchmarks associated with near-perfect agreement. Kolmogorov–Smirnov tests revealed statistically significant differences in the distributions of confidence values for most items, but the observed effect sizes were small, suggesting only modest divergence in how the two methods segment the confidence continuum. Mann–Whitney U tests found no significant difference in median confidence values for four items and only modest, consistently directional shifts for the others. In each case where a difference was detected, the relative method flagged responses with slightly higher confidence values than the absolute method—an expected outcome given its within-batch operational logic. These findings suggest that the relative method approximates the behavior of an absolute thresholding strategy not only in terms of response-level agreement but also in the distributional and central tendencies of flagged confidence values. The minor and systematic nature of these shifts underscores the method's practical validity, even in the absence of global thresholds.

One strength of the study is its multi-method evaluation strategy. By employing three distinct statistical tests—each aligned to a specific research question—the analysis provides a comprehensive and nuanced view of how the relative method compares to the absolute approach. This triangulation enhances the internal validity of the findings by ensuring that observed patterns are not artifacts of a single analytic lens. Prior research in assessment and machine scoring has emphasized the importance of using multiple indicators of agreement and reliability when evaluating human-machine alignment (e.g., Williamson, Xi, & Breyer, 2012). Extending this principle to thresholding methods strengthens the interpretive clarity of the current study.

A related strength is the use of operational data across ten unique items, which increases the generalizability of findings within the context of a real-world scoring system. Rather than relying on a narrow test set or simulated data, this study reflects real-world scoring dynamics, where batch effects, prompt variability, and distributional shifts routinely occur. Literature in automated writing evaluation has frequently called for validation

studies using authentic operational data (e.g., Bejar, 2011), and this study responds directly to that need.

Despite these strengths, a notable limitation is that the study treats the absolute threshold as a benchmark without fully interrogating its own validity or optimality. While the absolute method offers a theoretically attractive ideal—especially under conditions of complete data availability—it is not immune to its own biases, such as those introduced by non-uniform response distributions or scoring model calibration. A more complete validation strategy might compare both methods not only to each other but also to an external criterion, such as expert judgment of borderline cases.

Another area for future exploration involves the operational consequences of the observed differences. While McNemar's tests found no systematic disagreement in flagging decisions, statistical significance was more common in the comparisons of flagged-response distributions and central tendencies. The practical impact of these differences remains unclear. For example, do differences in flagged responses influence rater workload, response turnaround time, or score stability at the aggregate level? Future studies could simulate or analyze batch-level scoring flow under different flagging schemes to evaluate the impact of relative versus absolute methods on scoring efficiency and quality control. In this way, we might move beyond verifying that the relative method is good enough and begin to explore whether it is, in some cases, better suited to the realities of large-scale assessment.

In sum, the relative thresholding method performs robustly when compared to an absolute alternative, even though it makes decisions based only on within-batch information. It offers a stable, scalable solution that aligns well with theoretical expectations and empirical behavior of scoring confidence.

References

Bejar, I.I. (2011). A validity-based approach to quality control and assurance of automated scoring. Assessment in Education: Principles, Policy & Practice, 18(3), 319–341.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

McGraw, K.O., & Wong, S.P. (1992). A common language effect size statistic. Psychological Bulletin, 111(2), 361–365.

Vargha, A., & Delaney, H.D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics, 25(2), 101–132.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.