

Beyond Agreement: Operational Monitoring of K–12 Automated Scoring

Jing Ma¹, Edward W. Wolfe², Anthony Fina¹

¹Iowa Testing Programs

²The University of Iowa

Abstract

This study compares automated and human scores with expert/backread scores for writing, reading, and science tasks in a K-12 assessment program. Analyses examine discrepancies by score level and reused prompts across administrations. Automated scores showed smaller discrepancies for writing but larger discrepancies for short-answer tasks; reused prompts revealed localized shifts.

1 Background

Automated scoring (AS) is increasingly used in K-12 statewide assessment programs to reduce the time and cost associated with human scoring. A recent jurisdictional scan of U.S. state and Canadian provincial practices documents its growing operational use for constructed-response tasks, including short-answer items and essays. All U.S. states identified in the scan as using AS relied on commercial vendors for scoring services, with reported systems including the Intelligent Essay Assessor (IEA) and Project Essay Grade (PEG) (Zhang et al., 2026). Because scores from statewide summative assessments may inform high-stakes interpretations and decisions, the use of these systems requires evidence supporting both score reliability and the validity of the intended score interpretations and uses.

Evidence for AS must address both score reliability and validity of score interpretations and uses. Agreement indices such as quadratically weighted kappa, exact and adjacent agreement, and classification consistency are useful summaries of AS human correspondence, but they do not fully address whether scoring discrepancies are concentrated at particular score levels, task types, or administrations. Attali (2013) noted that automated human agreement for a single response does not capture variability across tasks, and Williamson et al. (2012) emphasized that validity evidence for

AS should extend beyond agreement to include construct representation, generalizability, fairness, and consequences. This concern is especially relevant for short-answer scoring, where responses are brief, content dependent, and often require fine distinctions among adjacent score categories. The present study does not attempt to explain the linguistic or cognitive mechanisms that make particular responses difficult to score. Instead, it uses this known scoring challenge as motivation for an operational monitoring approach that examines where AS and human scores differ from expert/backread scores across task types and score levels.

This rationale led to three design choices. First, operational evaluations commonly treat human scores as the criterion for evaluating AS, even though human scoring is also subject to error (Bennett and Bejar, 1997; Rotou and Rupp, 2020). When AS and human scores disagree, agreement statistics alone cannot indicate which score is closer to the intended scoring standard. Comparing both scores with an expert/backread score provides a common reference without assuming that operational human scores are error-free. Second, AS discrepancies may differ by task type and score level. This is important because writing essays and short-answer items differ in response length, construct representation, and rubric structure, and aggregate agreement indices can conceal localized over- or under-scoring. Third, operational monitoring requires evidence about whether AS performance remains stable across administrations (Rotou and Rupp, 2020). Writing prompts reused across years provide a practical opportunity to examine observed AS agreement and discrepancy patterns while holding prompt content constant. Together, these concerns motivate analyses that compare AS and human scores against expert/backread scores, examine conditional discrepancy patterns across task types and score levels, and evaluate reused prompts as a repeated prompt monitoring

comparison across administrations.

2 Aims

This study uses expert/backread scores as a common reference for comparing automated and operational human scores and examines reused writing prompts as a possible reference for operational monitoring. RQ1 asks how automated and human scores differ from expert/backread scores across expert score levels and task types; and RQ2 examines whether reused writing prompts can support operational monitoring by identifying whether observed AS agreement/discrepancy patterns differ across administrations.

3 Samples

Data came from a K–12 statewide assessment program and included open-ended responses scored by an automated scoring (AS) engine, operational human raters, and expert/backread raters. AS and operational human scores were generated through separate scoring processes and were treated as distinct scoring sources in the analysis. Expert/backread scores were assigned by specially trained senior raters through a separate backread review process and were used as the reference scores against which AS and operational human scores were evaluated.

The main cross-task analytic sample included only score units with all three scores available: AS, operational human, and backread. The analytic unit was the scored response for reading and science, and the domain-score unit for writing. Reading and science short-answer responses were scored on a 0–2 scale and contributed one score unit per response. Writing essays were scored on four domains, each on a 1–5 scale, so each essay response contributed four domain-score units, which are: Knowledge of Language (KOL), Text Types and Purposes (TTP), Production and Distribution of Writing (PDW), and Research to Build and Present Knowledge (RPK).

RQ1 sample focused on grades 5, 8, and 10 across four test administrations. Table 1 summarizes the number of items or prompts, responses, and scored units for each task type: reading short answer, science short answer, and writing essay. Because each task type corresponded to a different subject area in this operational dataset, task type differences should be interpreted as combined task type/subject area differences. The analytic sample included 6,359 reading short-answer score

units, 25,629 science short-answer score units, and 50,392 writing essay domain-score units from 12,598 essay responses. Because the tasks used different score scales, rater–expert differences were normalized by the score range for cross-task comparison.

Table 1: Main Cross-Task Analytic Sample

Item Type	Short Answer		Essay
Subject	Reading	Science	Writing
Items/Prompts	12	26	12
Scored Units	6,359	25,629	50,392
Responses	6,359	25,629	12,598

RQ2 focused on the Language/Writing test essay task. In the 2025 administration, one essay prompt in each grade was reused from an earlier administration between 2021 and 2023. This design provided an opportunity to compare AS performance across administrations for the same essay prompt. Table 2 shows the number of writing domain-score units available for each reused prompt in the prior administrations and the 2025 administration.

Table 2: Reused Writing Prompt Sample by Grade

Grade	Prior	2025
3	1,612	4,104
4	2,100	3,280
5	1,832	5,588
6	2,120	2,428
7	788	1,416
8	3,796	2,812
9	900	2,172
10	5,668	2,896
11	1,912	2,552

Note. Each grade represents one reused writing prompt that appeared in both the 2025 administration and one prior administration. Counts are writing domain-score units.

4 Methods

For RQ1, discrepancy indices were computed separately for AS and operational human scores using the expert/backread score as the reference. Let S_{ir} represent the score for score unit i from rater r , where r is AS or human, and let B_i represent the expert/backread score. The normalized directional difference was computed as

$$\text{NormDiff}_{ir} = \frac{S_{ir} - B_i}{\text{Score Range}}.$$

The normalized absolute difference was computed as

$$\text{NormAbsDiff}_{ir} = \frac{|S_{ir} - B_i|}{\text{Score Range}}.$$

Exact agreement was coded as 1 when $S_{ir} = B_i$ and 0 otherwise. Because exact agreement is affected by the number of score categories, exact agreement rates were interpreted primarily within each task type when comparing AS and operational human scoring. Cross-task comparisons emphasized normalized directional and absolute discrepancy indices. In this study, conditional analyses refer to evaluating each agreement and discrepancy measure at each expert/backread score level.

For reading and science short-answer items, the score range was 2. For writing essays, the score range was 4. These normalized indices allowed discrepancies to be compared across tasks with different score scales. Mean directional difference, mean normalized absolute difference, and exact agreement rate were reported overall by task type and rater and then by expert/backread score level. Within each score level cell, AS-only and Human-only error rates were also computed to identify whether discrepancies were more specific to AS or operational human scoring.

Generalized Estimating Equations (GEE) models were applied at the student-test level to account for dependence among score units from the same student. Separate models were estimated for normalized directional difference and normalized absolute difference. Both models examine rater, task type, normalized expert score position, and their interactions. Expert score position was normalized to a common 0–1 scale:

$$\text{BackPos}_i = \frac{B_i - \text{Minimum Score}}{\text{Score Range}}.$$

Both linear and quadratic expert score-position terms were included to capture score-level discrepancy patterns across the score scale. Grade and administration year were included as covariates to adjust for differences across grade levels and test administrations. Exact agreement was reported descriptively but was not modeled in the cross-task GEE analyses because agreement statistics, including exact agreement, are affected by the number of score categories (Wang and Dorans, 2021).

RQ2 focused only on the writing essay task. In the 2025 administration, one essay prompt in each grade was reused from an earlier administration between 2021 and 2023. Because each reused prompt

had a corresponding earlier administration, this design provided common prompt material for comparing observed AS agreement/discrepancy patterns across administrations. The analysis used all available writing grades and was restricted to domain-score units with a complete set of three scores: AS, operational human, and expert/backread. Normalized absolute discrepancy was summarized by administration, rater source, and expert/backread score level to examine the overall score level pattern across reused prompts. Prompt score cells were then compared between the prior administrations and 2025 to identify localized changes in AS exact agreement and absolute discrepancy.

5 Results

Regarding RQ1, Table 3 summarized within task comparisons of AS and operational human scoring using normalized directional difference, normalized absolute difference, and exact agreement with expert/backread scores. For reading short-answer items, AS had lower exact agreement (76.7%) than human scoring (84.3%) and a larger normalized absolute difference (0.116 vs. 0.079). The same pattern was stronger for science short-answer items: AS exact agreement was 80.2% versus 94.7% for humans, and normalized absolute difference was 0.099 versus 0.027. Writing essays showed the reverse pattern: AS was closer to expert/backread scores than operational human scoring in exact agreement (88.3% vs. 76.8%) and normalized absolute difference (0.029 vs. 0.058). The direction of the difference depended on task type.

Table 3: Overall Exact Agreement and Normalized Discrepancy by Task Type and Rater

Item Type	Short Answer		Essay	
	Rater	Reading	Science	Writing
Norm. Dir. Diff.	AS	-0.001	-0.004	-0.002
	Human	-0.014	-0.002	0.003
Norm. Abs. Diff.	AS	0.116	0.099	0.029
	Human	0.079	0.027	0.058
Exact Agreement	AS	76.7%	80.2%	88.3%
	Human	84.3%	94.7%	76.8%

Note. Norm. Dir. Diff. = mean normalized directional difference from the expert/backread score. Norm. Abs. Diff. = mean normalized absolute difference from the expert/backread score.

After conditioning on expert score level, Figure 1 shows mean normalized directional differences by rater source, and task type. AS had positive mean differences at expert score 0 and negative mean differences at expert score 2, with human scoring

showing the same directional pattern more weakly. Science short-answer responses showed a similar directional pattern, but the AS–human difference in mean directional discrepancy was smaller than in reading. For writing essays, human under-scoring became more pronounced than AS under-scoring at the upper end of the score scale.

Figure 2 shows corresponding task specific differences in normalized absolute discrepancy. AS had larger absolute differences than human scoring at every expert score level for both reading and science. The largest separation occurred for science at expert score 1, where AS absolute difference was substantially larger than human absolute difference (0.174 vs. 0.042), even though mean normalized directional differences were near zero. For writing essays, human absolute differences exceeded AS absolute differences at every expert score level. Consistent with this pattern, human-only error rates exceeded AS-only error rates at every score level, with the largest gap at the highest score point (36.2% vs. 11.9%). Complete score level results for all three task types are reported in Appendix Table A1.

Supporting GEE models were estimated for the two normalized discrepancy outcomes: normalized directional difference and normalized absolute difference. For normalized directional difference, the rater by task type interaction was significant, $\chi^2(2) = 564.06, p < .0001$, indicating that the AS human difference in signed discrepancy varied across task types. The three-way interactions among rater, task type, and expert score position were also significant for both the linear score position term, $\chi^2(2) = 20.46, p < .0001$, and the quadratic score position term, $\chi^2(2) = 22.33, p < .0001$.

For normalized absolute difference, the rater by task type interaction was also significant, $\chi^2(2) = 686.34, p < .0001$. The corresponding three-way interactions were significant for both the linear score position term, $\chi^2(2) = 455.06, p < .0001$, and the quadratic score position term, $\chi^2(2) = 319.13, p < .0001$. These model results support the descriptive finding that AS human differences varied across task types and changed across the expert/backread score scale. Because of the large sample size, the practical interpretation emphasized the descriptive magnitudes and graphical patterns rather than statistical significance alone.

For RQ2, the overall reused prompt results showed that AS remained closer to expert/backread

scores than operational human scoring in both the prior administration and 2025. As shown in Figure 3, mean normalized absolute discrepancy for AS was lower than human scoring at each expert score level in both administrations. The overall score-level pattern also suggested that observed AS discrepancy patterns were generally stable across administrations: AS normalized absolute discrepancy in 2025 was similar to or lower than the corresponding earlier administrations at most score levels.

However, the prompt score level results showed that the overall pattern did not capture all changes. Some reused prompt-score cells showed clear declines in AS exact agreement from the corresponding earlier administration to 2025. Table 4 displays the 10 reused writing prompt score cells with the largest declines in AS exact agreement. The largest decline occurred for the Grade 8 prompt at expert score 5, where AS exact agreement decreased by 18.6 points, from 97.3% in the prior administration to 78.7% in 2025. This decline was also accompanied by a 0.046 increase in normalized AS absolute discrepancy. The Grade 9 prompt at expert score 1 showed a 15.7 point decrease in AS exact agreement, and the Grade 6 prompt at expert score 4 showed a 12.4 point decrease.

The declines were not evenly distributed across all reused prompts. Seven of the 10 largest declines came from the Grade 8 and Grade 9 prompts, and several of these declines appeared across multiple expert score levels within the same prompt. Thus, the reused prompt analysis changed the interpretation of the overall writing pattern: aggregate AS absolute discrepancy appeared stable or improved, but selected prompts showed localized declines in exact agreement in 2025.

6 Discussion and Conclusion

These findings suggest that AS evaluation can benefit from moving beyond omnibus agreement indices to examine discrepancies as a function of rater source, task type, and expert score level. Prior work has emphasized agreement, accuracy, and validity evidence for automated scoring systems (Bennett and Bejar, 1997; Williamson et al., 2012; Liu et al., 2014), and recent work shows that AS is increasingly used in K-12 summative assessment programs (Zhang et al., 2026). The present study adds to this work by comparing AS and operational human scores against a common expert/backread

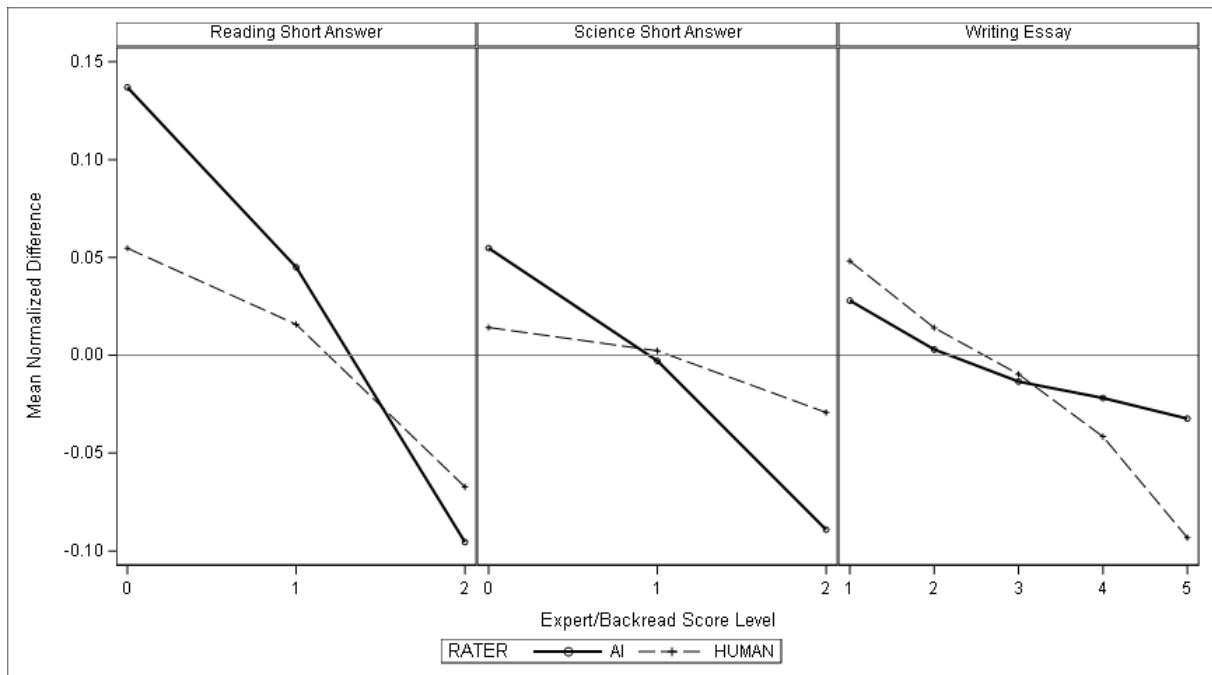


Figure 1: Mean Normalized Directional Difference by Expert/Backread Score Level, Task Type, and Rater

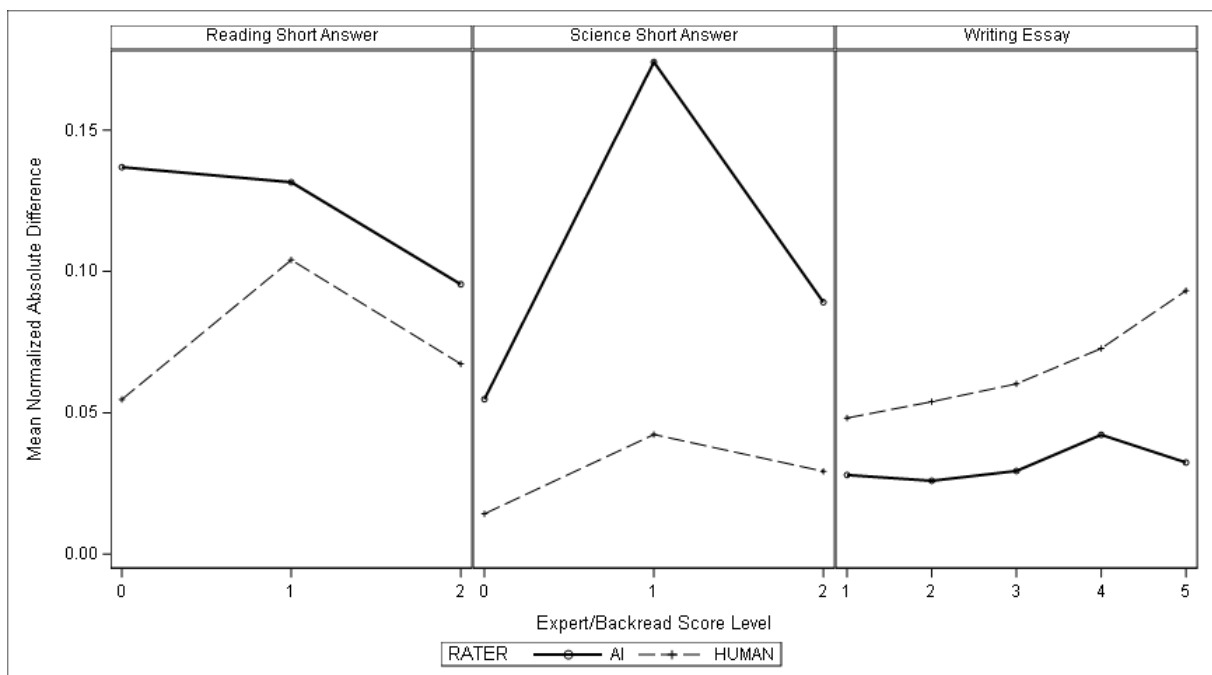


Figure 2: Mean Normalized Absolute Difference by Expert/Backread Score Level, Task Type, and Rater

benchmark and by examining where discrepancies occur.

The results show that scoring risk was not evenly distributed across the assessment. Reading and science short-answer tasks showed more AS specific discrepancy than writing essays. In reading, AS showed a score-level directional pattern, with over scoring at the lowest expert score level and under scoring at the highest score level, while AS abso-

lute discrepancy remained larger than human discrepancy across all reading score levels. Science also showed larger AS discrepancy than human scoring, with the clearest concern concentrated at the partial credit score level. Writing showed a different pattern: AS was generally closer to expert/backread scores than operational human scoring, especially at higher score levels.

These findings support a risk based approach

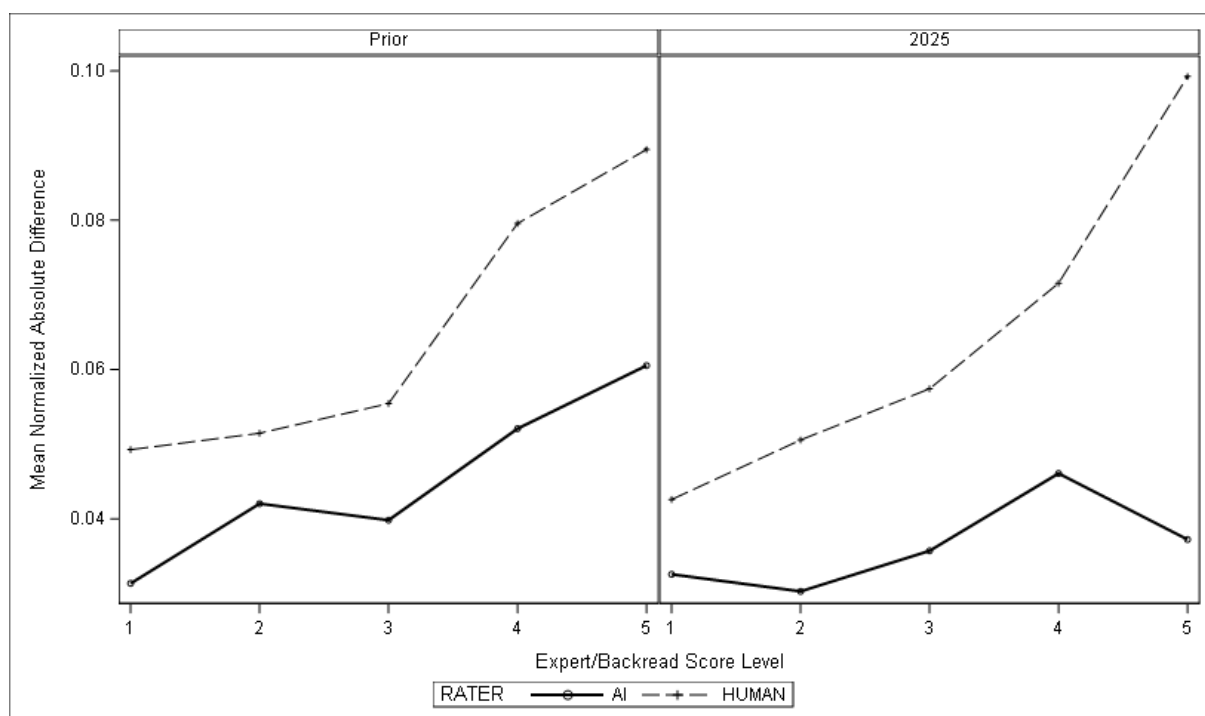


Figure 3: Prior versus 2025 absolute discrepancy by expert score for reused writing prompts.

Table 4: Reused Writing Prompt-Score Cells with Largest Declines in AS Exact Agreement

Grade	Expert Score	Prior <i>N</i>	2025 <i>N</i>	Prior AS Exact	2025 AS Exact	Change in AS Exact (points)	Change in AS Norm. Abs. Diff.
8	5	73	122	97.3%	78.7%	-18.6	0.046
9	1	146	331	95.2%	79.5%	-15.7	0.039
6	4	126	187	87.3%	74.9%	-12.4	0.029
9	4	66	191	97.0%	85.9%	-11.1	0.028
9	2	282	776	94.0%	83.5%	-10.5	0.026
8	3	1,277	1,059	93.0%	83.9%	-9.1	0.023
8	4	555	657	94.2%	85.4%	-8.8	0.022
8	2	1,257	704	89.7%	80.8%	-8.8	0.022
11	5	63	33	90.5%	81.8%	-8.7	0.022
11	4	318	315	86.5%	78.4%	-8.1	0.020

Note. Prior refers to the corresponding earlier administration for each reused prompt. *N* = writing domain-score units. Prior AS Exact and 2025 AS Exact are the percentages of domain-score units for which the AS score matched the expert/backread score. Change in AS Exact = 2025 AS Exact minus Prior AS Exact, reported in percentage points. Change in AS Norm. Abs. Diff. = 2025 normalized AS absolute difference minus prior normalized AS absolute difference.

to AS monitoring. Rather than applying the same level of review across all tasks and score levels, programs can use conditional discrepancy evidence to identify where closer review may be most useful. For short-answer tasks, this may include additional backreading, scorer review, rubric clarification, or targeted AS retraining in score regions where discrepancies are concentrated. For writing essays, monitoring may need to focus more on prompt specific changes over time, rubric interpretation, and consistency between operational human scoring and expert/backread expectations.

The reused prompt analysis offers one practical

monitoring strategy. Because the same prompts appeared in different administrations, they provided common prompt material for examining observed AS agreement/discrepancy patterns when prompt content was held constant. The results showed that overall AS agreement/discrepancy patterns appeared stable or improved, but some reused prompts showed declines in exact agreement with expert/backread scoring at certain score levels. These findings suggest that reused prompts can help detect score level weak spots that are hidden in aggregate results. At the same time, year-to-year comparisons should be interpreted cautiously

because observed differences may reflect not only prompt-specific scoring behavior, but also changes in the AS system, scoring procedures, or operational context across administrations.

This study has several limitations. Although operational human and expert/backread records were generated through separate scoring processes, the analysis did not separately evaluate how additional score resolution rules shaped the available scoring records. Expert/backread scores provided a common benchmark but should not be interpreted as error-free. The reused prompt comparisons were descriptive and may reflect differences in response distributions, operational conditions, scorer pools, or AS system configuration across administrations. Some prompt by score cells were small, so localized changes require confirmation with additional data. In addition, task type and subject area were aligned in this dataset: reading short answer, science short answer, and writing essay represented both different subject areas and different response formats. Therefore, the observed task type differences cannot be attributed only to response format.

Future work should take a closer look at the prompts and scoring rubrics associated with the largest discrepancies to identify possible sources of scoring difficulty and potential improvements to AS training, rubric clarification, or scoring procedures.

References

- Y. Attali. 2013. Validity and reliability of automated essay scoring. In M. D. Shermis and J. Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pages 181–198. Routledge/Taylor & Francis Group.
- Randy Elliot Bennett and Isaac I. Bejar. 1997. [Validity and automated scoring: It's not only the scoring](#). *ETS Research Report Series*, 1997(1):i–30.
- Ou Lydia Liu, Chris Brew, John Blackmore, Libby Gerard, Jai Madhok, and Marcia C. Linn. 2014. [Automated scoring of constructed-response science items: Prospects and obstacles](#). *Educational Measurement: Issues and Practice*, 33(2):19–28.
- Ourania Rotou and André A. Rupp. 2020. [Evaluations of automated scoring systems in practice](#). *ETS Research Report Series*, 2020(1):1–18.
- W. Wang and N. J. Dorans. 2021. [Impact of categorization and scaling on classification agreement and prediction accuracy statistics](#). *ETS Research Report Series*, 2021:1–20.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.
- S. Zhang, J. Hove, and Z. Gu. 2026. A jurisdictional scan of automated scoring practices in K–12 summative assessment programs. Webinar. State and Local Assessment Leaders, August 6, 2026.

Appendix

A Detailed Score Level Results

Table A1: Agreement and Discrepancy by Task Type and Expert/Backread Score

Task Type	Expert/ Backread Score	<i>N</i>	AS Abs. Diff.	Human Abs. Diff.	AS Exact	Human Exact	AS-only Error	Human-only Error
Reading Short Answer	0	1,161	0.137	0.055	72.6%	89.1%	27.3%	10.9%
	1	2,344	0.132	0.104	73.7%	79.2%	26.3%	20.7%
	2	2,854	0.095	0.067	80.9%	86.5%	19.1%	13.4%
Science Short Answer	0	10,698	0.055	0.014	89.1%	97.2%	10.6%	2.6%
	1	7,426	0.174	0.042	65.2%	91.5%	32.8%	6.4%
	2	7,505	0.089	0.029	82.3%	94.2%	17.0%	5.1%
Writing Essay	1	7,558	0.028	0.048	89.0%	80.9%	9.6%	17.7%
	2	19,804	0.026	0.054	89.7%	78.4%	9.5%	20.7%
	3	16,419	0.029	0.060	88.3%	75.9%	10.6%	23.0%
	4	5,584	0.042	0.073	83.1%	71.0%	15.2%	27.4%
	5	1,027	0.032	0.093	87.1%	62.8%	11.9%	36.2%

Note. AS denotes automated scoring. Absolute differences were normalized by the possible score range. Exact agreement is the percentage of scores equal to the expert/backread score. AS-only errors occurred when AS disagreed with the expert/backread score and human scoring did not; Human-only errors indicate the reverse.