Peer Beyond Graduate Research Journal

Peer Beyond Graduate Research Conference | Volume 2, Issue 1| Pages 39–42

Reliability by Design: Human–AI Writing Assessment

Ali Mikaeili Barouq
University of Calgary.

Email: ali.mikaeilibarouq@ucalgary.ca

Abstract: High-stakes writing assessment is a standards-referenced judgment practice in which reliability depends on shared interpretations of quality rather than mechanical measurement. This study examines reliability and standards communication in a rubric-driven, AI-supported system for Alberta English Language Arts 30–1. Fifteen publicly released exemplar essays (2022–2024) were evaluated using verbatim rubric embedding, criterion-level evidence requirements, and a gated decision process without numerical aggregation. Reliability was assessed through alignment with authorized classifications and stability across repeated evaluations (150 decisions). The system matched official classifications in 93.3% of cases and was fully stable across repetitions, with all disagreements confined to adjacent performance boundaries. Qualitative comparison of teacher and AI commentaries indicates that the system more consistently externalized criterion-referenced warrants, enhancing transparency without displacing human authority. Findings frame reliability as an emergent property of assessment-system design and position AI as infrastructure for contestable standards communication.

Introduction

Assessment is not only a measurement problem; it is a sensemaking problem. In standards-referenced writing assessment, quality is determined through qualitative comparison to articulated standards, with expert judgment doing the heavy lifting (Sadler, 1987, 1989). Rubrics do not automate correctness; they mediate interpretation by specifying what counts as quality and how performance distinctions are recognized (Dawson, 2017). Yet even with rubrics, judgments can diverge—especially at boundaries where student work partially satisfies competing descriptors (Kuiken & Vedder, 2014). From a validity perspective, the key question is not whether scores match, but whether interpretations and uses of results are warranted and defensible (Kane, 2013).

Large language models (LLMs) have renewed interest in automated scoring and feedback. Studies report that LLM-based scoring can approach conventional agreement benchmarks under controlled prompting (Shermis, 2025) and that intra-rater stability can be strong in constrained settings (Pack et al., 2024). At the same time, concerns persist about construct representation and the tendency of automated systems to rely on proxies (e.g., surface linguistic features) rather than rubric-aligned qualities of reasoning and rhetorical effectiveness (Kumar & Boulanger, 2021). For learning sciences audiences, a productive shift is to treat AI not as a replacement rater, but as a tool that can make the enactment of standards more visible, auditable, and discussable—supporting professional learning and contestability while preserving human authority (Aloisi, 2023; Richardson & Clesham, 2021).

Empirical research complicates assumptions that rubric alignment produces interchangeable judgments between humans and AI systems. Comparative studies report low to moderate agreement even under shared analytic rubrics and calibration protocols, indicating that rubrics constrain interpretation without standardizing it (Wetzler et al., 2025). Divergences are patterned rather than random: AI systems demonstrate proportional bias across performance levels and greater stability for surface linguistic features than for higher-order rhetorical and conceptual dimensions (Bui & Barrot, 2025; Zhao et al., 2023). Agreement indices alone therefore provide an incomplete account of reliability. Although many-facet Rasch models estimate rater severity and task effects (Yamashita, 2024), high internal consistency can coexist with systematic interpretive misalignment (Jin et al., 2025), and short-term stability may deteriorate through model drift (Hackl et al., 2023). Beyond agreement, concerns about construct representation persist. Deep learning–based systems may exceed human–human agreement while relying on opaque feature proxies that lack transparent alignment with rubric-defined constructs (Kumar & Boulanger, 2020, 2021). Studies of rubric explicitness further demonstrate that stability improves only when evaluative logic is externalized through detailed descriptors and decision rules (García-Varela et al., 2025; Wu et al., 2025). Taken together, this literature suggests that reliability in AI-assisted writing assessment is less a property of models than an outcome of assessment-system design and standards mediation (Farzi, 2024; Bucol & Sangkawong, 2025).

This study tests that design stance in a high-stakes, standards-referenced context: Alberta ELA 30–1. We evaluate whether a rubric-driven, evidence-required AI system can (a) produce stable performance-level recommendations and (b) communicate standards through transparent, criterion-referenced warrants.

Research Questions

RQ1. How consistent are AI-supported performance-level recommendations compared with authorized human performance classifications in a high-stakes writing assessment context?

RQ2. How do AI-generated commentaries justify judgments in relation to rubric criteria and standards?

RQ3. How do human and AI commentaries converge or diverge in reasoning at adjacent performance boundaries?

Methodology

Fifteen publicly released Alberta ELA 30–1 exemplar essays from 2022–2024 were selected when the official performance level and accompanying commentary were available. Official classifications were treated as authorized reference judgments for system-to-reference comparison (Alberta Education, 2022, 2023, 2024).

The system was built as a rubric-mediated judgment procedure rather than a generative grader. The official ELA 30–1 scoring guide was embedded verbatim, with no added criteria, numeric weights, or composite scoring. For each essay, the system produced criterion-level judgments aligned to the relevant rubric categories, required textual evidence (quotations or explicit text references) before stating evaluative claims, and produced an overall performance level via a gated, exclusion-based procedure (e.g., a single criterion pattern excluding a higher level prevented an upgrade). This design follows standards-referenced logic: thresholds are qualitative and defensible when the evidence–standard link is explicit (Sadler, 1987, 1989; Kane, 2013).

System reliability was examined in two ways: alignment (exact agreement between system recommendations and authorized performance levels across ordered categories) and intra-system stability (each essay evaluated 10 times under identical conditions; $15 \times 10 = 150$ decisions). Standards communication was examined qualitatively by comparing teacher and AI commentaries as artifacts of judgment. Drawing on criteria-referenced feedback principles, attention focused on whether commentaries (i) explicitly referenced criteria/standards, (ii) grounded claims in textual evidence, and (iii) made the warrant linking evidence to level visible (Brookhart, 2017).

Results

Across 150 repeated decisions, the system's performance-level recommendations were 100% stable under identical conditions (no classification changes across runs). For system-to-reference alignment, the system matched authorized classifications in 140/150 decisions (93.3%). All ten mismatches occurred at adjacent boundaries (Satisfactory–Proficient or Proficient–Excellent); there were no non-adjacent disagreements.

Inspection of mismatch cases indicated a consistent pattern: both sources (teacher and system) typically cited overlapping textual evidence and operated within the same rubric categories but differed in how they weighted evidence relative to threshold descriptors. In learning sciences terms, disagreements concentrated where the task demands interpretive calibration—precisely the zone where standards-referenced assessment predicts judgment pressure to be highest (Kuiken & Vedder, 2014; Sadler, 1987).

Teacher commentaries often condensed reasoning into professional shorthand, while AI commentaries more consistently externalized warrants: explicit criterion reference, direct evidence citation, and why this meets/does not meet the descriptor explanations. This does not imply superior judgment; it reflects the communicative affordance of an evidence-required design. In several boundary cases, the AI's transparency made the disagreement legible: readers could identify the exact threshold claim (e.g., whether conceptual integration was thoughtful versus insightful) and contest it using the same rubric language.

Discussion

For learning sciences, the headline is not AI can grade. It is that reliability can be engineered as a design outcome when standards are embedded verbatim, applied criterion-first, and paired with evidence-required warrants. This reframes reliability from a trait of raters or models to an emergent property of an assessment system—a socio-technical arrangement that shapes how judgments are produced, inspected, and discussed (Sadler, 1987, 1989; Kane, 2013).

High alignment plus perfect intra-system stability under fixed conditions suggests that design constraints can reduce procedural noise and make performance recommendations reproducible. This complements LLM scoring findings that emphasize controlled prompting for stability (Pack et al., 2024; Shermis, 2025) but shifts the focal unit from the model to the procedure.

The confinement of disagreements to adjacent boundaries suggests threshold sensitivity rather than categorical failure. In standards-referenced writing assessment, boundary decisions are inherently interpretive; disagreement is often a feature of judgment work, not merely error (Sadler, 1987). A system that preserves ordinality and concentrates divergence at boundaries is behaving more like a standards-referenced assessor than a proxy-feature calculator.

The strongest contribution may be standards communication. When AI outputs consistently surface criterion-referenced warrants, they can function as infrastructure for transparency and contestability—supporting moderation, professional dialogue, and learner-facing explanation—while keeping consequential authority with humans (Brookhart, 2017; Aloisi, 2023; Richardson & Clesham, 2021). This matters for legitimacy: in high-stakes contexts, trust depends not only on outcomes but on whether reasoning trails are inspectable and challengeable.

Conclusion

A rubric-driven, evidence-required human–AI assessment system achieved 93.3% agreement with authorized high-stakes performance classifications and 100% stability across repeated identical evaluations, with all disagreements confined to adjacent performance boundaries. Qualitative comparison suggests that AI can strengthen standards communication by externalizing criterion-referenced warrants, enabling more transparent and contestable judgment without displacing human authority. For learning sciences, the results support a design-oriented claim: in standards-referenced writing assessment, reliability is best pursued as a property of assessment-system design, not as a promise of automated replacement.

References

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education, 58*(1), 98–110. https://doi.org/10.1111/ejed.12542

Alberta Education. (2022). *English Language Arts 30–1 scoring guide: Scoring categories and scoring criteria for 2024–2025.* Government of Alberta.

Alberta Education. (2023). *English Language Arts 30–1: Examples of the standards for students' writing from the January 2023 diploma examination.* Government of Alberta.

Alberta Education. (2023). *English Language Arts 30–1: Examples of the standards for students' writing from the June 2022 diploma examination.* Government of Alberta.

Alberta Education. (2024). *English Language Arts 30–1: Examples of the standards for students' writing from the January 2024 diploma examination.* Government of Alberta.

Brookhart, S. M. (2017). *How to give effective feedback to your students* (2nd ed.). ASCD.

Bucol, J. L., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International, 62*(3), 867–882. https://doi.org/10.1080/14703297.2024.236390

Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*, 1–16. https://doi.org/10.1080/17501229.2025.2521003

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education, 42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Farzi, R. (2024). Calibrating generative AI for second language writing assessment: Combining statistical validation with prompt design. *Assessment and Practice in Educational Sciences, 2*(4), 1–12. https://www.journalapes.com/index.php/apes/article/view/91

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences, 15*(8), Article 946. https://doi.org/10.3390/educsci15080946

Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education, 8*, Article 1272229. https://doi.org/10.3389/feduc.2023.1272229

Jin, R., Zhao, M., Niu, C., Xia, Y., Zhou, H., & Liu, N. (2025). Evaluating the performance of ChatGPT and Claude in automated writing scoring: Insights from the many-facet Rasch model. *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13774-4

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing, 31*(3), 279–284. https://doi.org/10.1177/0265532214526179

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education, 5*, Article 572367. https://doi.org/10.3389/feduc.2020.572367

Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education, 31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of AI technologies in high-stakes assessment. *London Review of Education, 19*(1), Article 9. https://doi.org/10.14324/LRE.19.1.09

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209. https://doi.org/10.1080/0305498870130207

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

Shermis, M. D. (2025). Using ChatGPT to score essays and short-form constructed responses. *Assessing Writing, 66*, Article 100988. https://doi.org/10.1016/j.asw.2025.100988

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2025). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology, 52*(3), 298–304. https://doi.org/10.1177/00986283241282696

Wu, X., Saraf, P. P., Lee, G., Latif, E., Liu, N., & Zhai, X. (2025). Unveiling scoring processes: Dissecting the

differences between LLMs and human graders in automatic scoring. *Technology, Knowledge and Learning*, Article 100177. https://doi.org/10.1007/s10758-025-09836-8

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics, 3*(3), Article 100133. https://doi.org/10.1016/j.rmal.2024.100133

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies, 28*(6), 7031–7063. https://doi.org/10.1007/s10639-022-11473-y