

Beyond Agreement: Calibrating Automated Scoring to Human Judgment Using Control Scripts

Mark Dulhunty, Alex Codoreanu, and Nathan Zoanetti

Australian Council for Educational Research
Melbourne, Victoria, Australia

Correspondence: mark.dulhunty@acer.org

Abstract

Automated scoring systems are typically evaluated using agreement with human-assigned scores, yet deployment in operational assessment requires evidence of calibration, reproducibility, and alignment with existing quality assurance processes. This study evaluates an ensemble automated scoring approach across 55 constructed-response reading items using approximately 260,000 student responses. Initial large-scale evaluation established strong agreement with human-assigned scores. Model performance was subsequently calibrated against operational reference scores using 477 embedded control scripts and compared with 27 human markers. The ensemble achieved 95.3% agreement with operational reference scores, matching or exceeding individual human marker performance on many items. Reproducibility testing demonstrated complete scoring stability across repeated evaluations, providing evidence that the approach generated consistent outcomes under operational conditions. High-performing models were then applied within marking workflows to support quality assurance through targeted review of disagreements between automated and human-assigned scores. The findings demonstrate how automated scoring can be evaluated and calibrated within operational assessment systems, supporting its use as an embedded measurement and quality assurance function rather than solely a predictive tool.

Melconian, 2026; Rotou & Rupp, 2020). Much existing research, including recent large language model and ensemble-based approaches (e.g., Frohn et al., 2025a; Xue et al., 2026), treats automated scoring as a predictive task, emphasising accuracy relative to reference scores (e.g., Ahmed et al., 2022; Henkel et al., 2025; Jung et al., 2026). In operational contexts, however, it must function as a measurement instrument, requiring evaluation beyond accuracy to consider alignment with human judgement and quality assurance practices.

This study addresses this gap through a multi-phase, program-aligned evaluation, of automated scoring in practice. An initial large-scale evaluation was conducted to establish model performance relative to human markers ($N \approx 260,000$ responses), providing evidence of reliability at scale. Building on this, control scripts (i.e., responses with established reference scores) were embedded within operational marking. Throughout this paper, reference scores are treated as the operational scoring standard. These scores were established through lead-marker consensus and should be interpreted as operationally validated rather than objectively 'true' scores (see Sample section for details). These scripts were scored by both human markers and automated models, enabling evaluation against reference standards and direct comparison with individual human marker performance. In summary, the study follows a three-phase evaluation logic: large-scale evaluation to establish reliability, control script evaluation against reference scores, and operational deployment to assess practical utility.

1 Introduction

Evaluating automated scoring in operational assessment requires situating model performance relative to both scoring standards and observed human marker behaviour (Frohn et al., 2025b;

2 Aims

This study aims to:

1. Evaluate agreement of the ensemble automated scoring approach relative to

human scoring at scale across constructed-response items.

2. Assess calibration through comparison with reference scores and individual human marker performance on control scripts.
3. Examine the extent to which the model can be applied to operational responses to support identification of inconsistent human judgements.

3 Sample(s)

The study draws on multi-phase operational assessment data from a large-scale reading program. The primary evaluation dataset comprises 55 dichotomously scored constructed-response items and approximately 260,000 student responses. Responses were typically short (ranging from phrases to one or two sentences) such as *“There is a supermarket that takes up a lot of room”*.

Control scripts ($N = 477$) with established reference scores were embedded within operational marking and scored by at least two lead markers, with individual markers ($N = 27$) scoring varying subsets of scripts due to operational allocation constraints. The number of control scripts varied by item (ranging from 4 to 14 control scripts depending on the item).

A subset of items ($N=10$), for which the ensemble automated scoring approach demonstrated strong performance, was selected for a later operational phase ($N \approx 14,000$ responses) to evaluate its application for quality assurance.

4 Methods

The study follows three phases: large-scale evaluation to establish reliability, validation against control scripts with known scores, and operational deployment to assess practical utility.

Control scripts with established reference scores were embedded within operational marking, allowing human and automated scoring of shared responses under operational conditions. This design enabled direct comparison of model outputs with individual human markers within a shared measurement framework.

Automated scoring combined generative and embedding-based approaches within an ensemble

framework (see Appendix A for full model specifications and Table A1). The ensemble integrated outputs from five model components differing in architecture, training data, and prompt configuration, with final scores assigned using a majority voting rule. Although components were not fully independent due to shared data sources, this design introduces controlled diversity, improving robustness and consistency (Xue et al., 2026).

Agreement estimates were calculated using approximately 240,000 held-out responses, with approximately 10% of available responses used for model training across the transformer based ensemble components. Model performance was evaluated against reference scores and relative to individual human marker performance on control scripts, supporting comparison against both scoring standards and observed variability in human marker behaviour. Additional agreement analyses were conducted on operational responses using human-assigned scores to assess broader performance beyond control scripts.

To assess reproducibility, a subset of approximately 50,000 responses was repeatedly scored using the same ensemble configuration. This analysis was conducted to evaluate score stability and address potential concerns that generative LLM components may produce different outputs when repeatedly presented with the same response. Repeated scoring outcomes were compared to assess the consistency of score assignments across runs.

Following this evaluation, a subset of items was selected for which the ensemble automated scoring approach demonstrated strong performance based on agreement and calibration relative to both reference scores and human marker performance on control scripts. Selection was further informed by consistently high agreement, defined as exceeding 95% agreement during initial evaluation across all individual scoring components. The model was then applied to operational responses for these items to generate disagreement reports during marking, comparing automated and human-assigned scores to identify responses for targeted review within existing quality assurance processes.

The overall evaluation and deployment framework is illustrated in Figure 1.

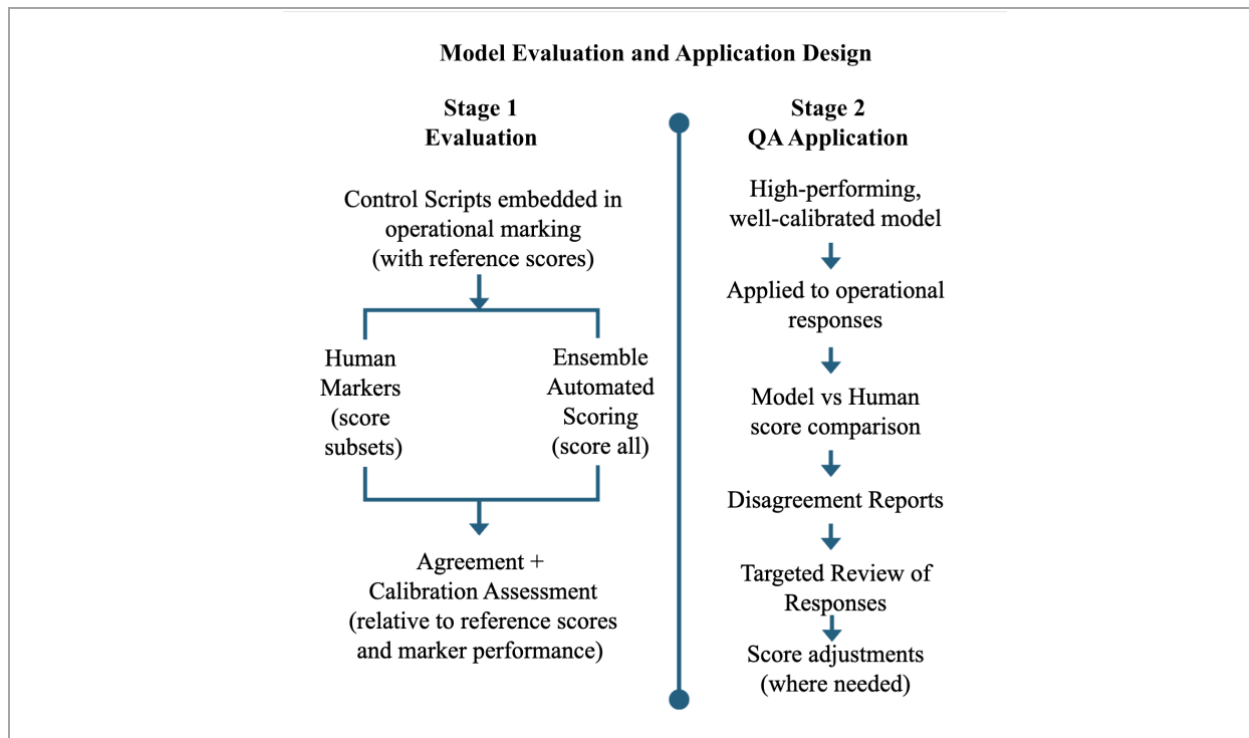


Figure 1: Following large-scale evaluation, embedded control scripts enable human-referenced evaluation of automated scoring (Stage 1), supporting deployment of calibrated models for quality assurance (QA) through disagreement detection (Stage 2).

5 Results

Predicted scores for each component of the ensemble automated scoring approach were initially evaluated against approximately 260,000 human-marked responses across 55 items. Each component achieved an average item-level Cohen's kappa (κ) greater than 0.70 (range: 0.73–0.78), indicating strong agreement with human markers and supporting progression to subsequent evaluation phases.

Evaluation of the final ensemble demonstrated similarly strong performance across most items (Figure 2). Of the 55 items evaluated, 44 (80%) achieved item-level κ values above 0.6, including 26 items (47%) above $\kappa = 0.8$. Only two items returned κ values below 0.4, with the majority of items demonstrating moderate to strong agreement.

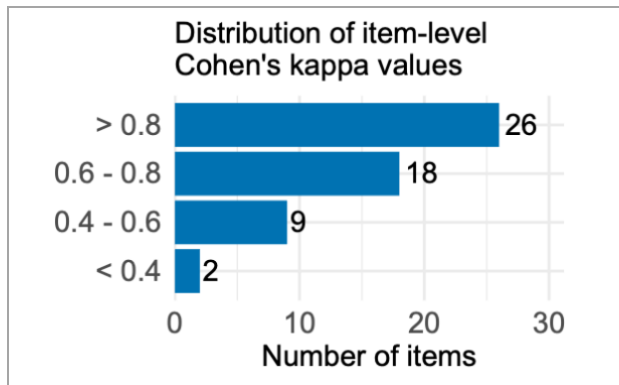


Figure 2: Distribution of item-level Cohen's κ values for the ensemble automated scoring approach across 55 reading constructed-response items.

The ensemble approach was then evaluated against 477 control scripts with established reference scores. Across these control scripts, the ensemble achieved 95.3% agreement with operational reference scores. In addition, perfect agreement was observed for 39 of 55 items (71%), although individual items contained between 4 and 14 control scripts. Performance on many items was comparable to or exceeded that of many individual human markers.

The ensemble ‘AutoScore’ performance against individual human marker performance on control scripts is shown in Figure 3. Figure 3 shows that AutoScore performed similarly to or better than most markers within the comparison pool.

In Figure 3, ‘Similar’ performance was when AutoScore was within $\pm 0.5\%$ of human marker accuracy. Comparisons were restricted to markers with substantial evidence available (≥ 21 items and

≥ 180 control scripts). AutoScore performed similarly to or better than 18 of the 22 markers in the comparison pool. AutoScore underperformed compared to 4 markers, who were also the top performing markers across the entire marking pool.

Direct comparison with individual marker performance showed that model scores were well calibrated relative to both reference standards and observed variability in human marking behaviour.

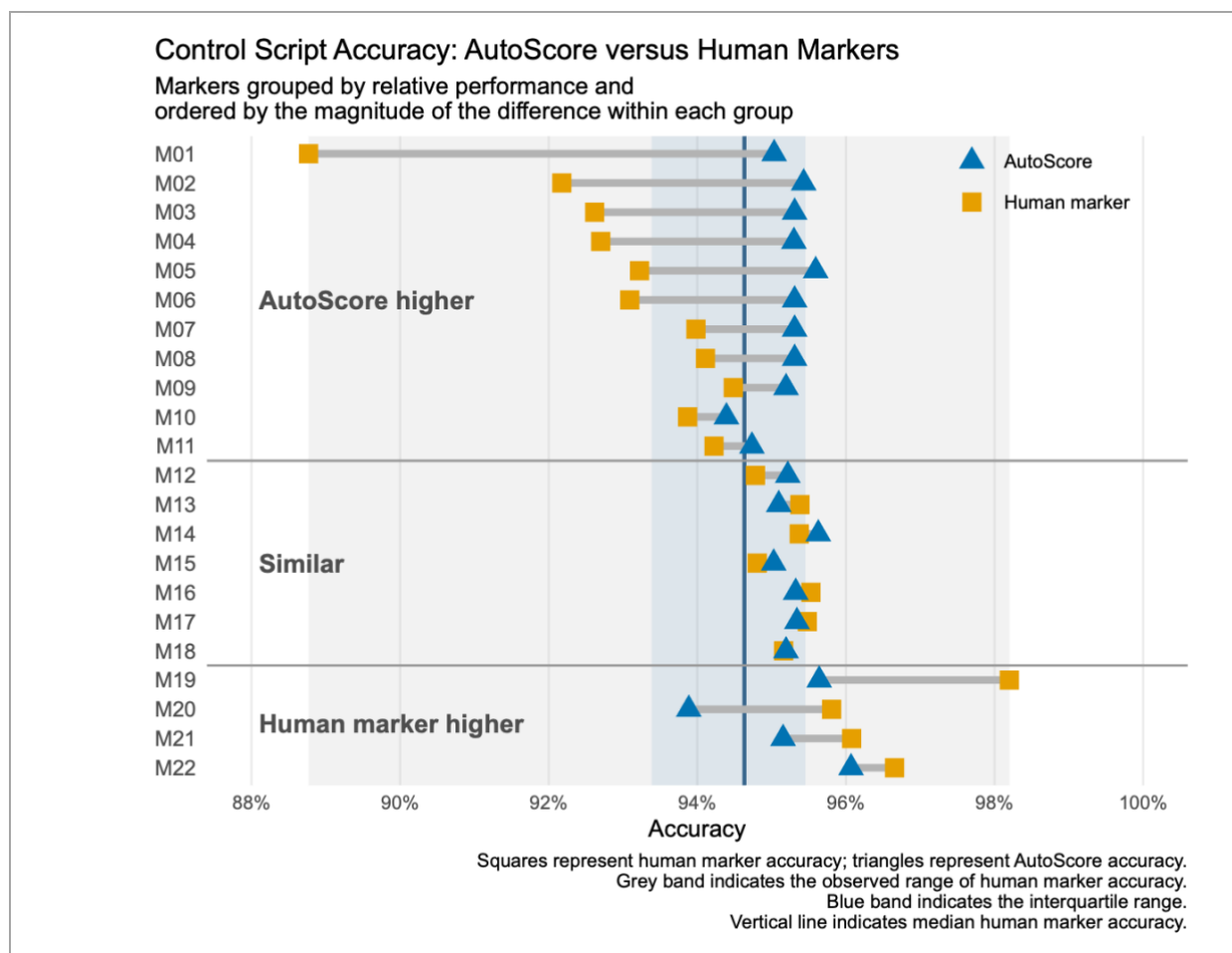


Figure 3: Comparison of AutoScore and human marker accuracy on control scripts. Markers are grouped according to whether AutoScore achieved higher, similar, or lower accuracy than the human marker. Detailed marker-level results are provided in Appendix B.

Reproducibility testing demonstrated complete scoring stability across repeated evaluations, with identical scores assigned when the same responses were evaluated repeatedly (based on a sample of approximately 50,000 responses).

When applied to operational responses, high-performing models identified a small subset of disagreements indicative of potential scoring inconsistencies. Although overall agreement was

high and the number of disagreements was correspondingly low, targeted review of these cases resulted in 13 confirmed score adjustments from approximately 14,000 operational responses, demonstrating the capacity of automated scoring to support existing quality assurance processes.

The proportion of disagreements was lower than typical quality assurance sampling thresholds, suggesting that the approach can support more

targeted and efficient review while maintaining confidence in scoring outcomes. Taken together, these findings indicate that well-calibrated automated scoring models can be deployed to support quality assurance processes through the targeted identification of potentially inconsistent human judgements, while preserving existing operational workflows.

6 Conclusions

Findings show that well-calibrated automated scoring models can achieve consistent performance across items, in some cases matching or exceeding individual human marker performance on control scripts while remaining aligned with established scoring standards. When integrated into operational workflows, such models can support quality assurance through targeted review of potentially inconsistent human judgements without disrupting existing processes.

Reliable operational scoring requires that identical responses receive identical scores when processed repeatedly. Reproducibility testing therefore complements agreement and calibration analyses by confirming that the automated scoring process behaves deterministically in practice. Results demonstrated complete scoring stability across repeated evaluations, showing that ensemble-based automated scoring can provide stable and repeatable scoring outcomes despite incorporating generative LLM components.

This study demonstrates that automated scoring can be evaluated and calibrated within operational assessment systems through comparison with both reference scores and observed human marker behaviour on control scripts. This human-referenced calibration framework provides a practical basis for interpreting model performance relative to operational scoring standards, extending evaluation beyond aggregate accuracy metrics alone.

More broadly, this work reframes automated scoring from a purely predictive task to an embedded measurement function, demonstrating how model evaluation, calibration, and deployment can be integrated within ongoing scoring and quality assurance practices.

This study forms part of a broader programme of research investigating automated scoring methods, infrastructure, and operational

deployment within large-scale assessment systems. Future work will extend the calibration framework presented here to additional item types, broader rubric-based scoring approaches, and other applications of AI-assisted assessment.

7 Limitations

This study has several limitations. First, the number of control scripts per item was relatively small (4-14), reflecting operational rather than experimental data collection. Second, control scripts were selected for quality assurance purposes rather than randomly sampled, which may limit generalisability. Third, individual markers completed differing subsets of control scripts because of operational allocation requirements, resulting in unequal evidence across marker-level comparisons. Finally, the study focused on dichotomously scored reading items within a single assessment program, and findings may not generalise to other item types or scoring contexts.

Acknowledgements

The authors gratefully acknowledge the support of ACER colleagues, particularly Lynn Sendy-Smithers, Jenna Schorel, and Rose Knight, whose contributions helped make this work possible.

References

- Ahmed, A., Joorabchi, A., & Hayes, M. J. (2022). On the application of sentence transformers to automatic short answer grading in blended assessment. In *2022 33rd Irish Signals and Systems Conference (ISSC 2022)*. Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ISSC55427.2022.9826194>
- Frohn, S., Burleigh, T., & Chen, J. (2025a). Automated scoring of short answer questions with large language models: Impacts of model, item, and rubric design. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education (AIED 2025)* (Lecture Notes in Computer Science, Vol. 15882). Springer. https://doi.org/10.1007/978-3-031-98465-5_6
- Frohn, S., Deters, L., Burleigh, T., & Dicerbo, K. (2025b). A framework for live scoring constructed response items with commercial LLMs. Paper presented at the Artificial Intelligence in Measurement and Education Conference, Pittsburgh, PA. https://www.researchgate.net/publication/396937943_A_Framework_for_Live_Scoring_Constructed_Response_Items_with_Commercial_LLMs

- Henkel, O., Hills, L., Roberts, B., et al. (2025). Can LLMs grade open response reading comprehension questions? An empirical study using the ROARs dataset. *International Journal of Artificial Intelligence in Education*, 35, 651–676. <https://doi.org/10.1007/s40593-024-00431-z>
- Jung, J. Y., Bezirhan, U., & von Davier, M. (2026). Optimizing automated scoring in ILSAs with prompt compression. *Computers & Education: Artificial Intelligence*. <https://doi.org/10.1016/j.caeai.2026.100558>
- Melconian, M. (2026, April 8). Inter-rater agreement and AI scoring: The reliability evidence you now need. LinkedIn. <https://www.linkedin.com/pulse/inter-rater-agreement-ai-scoring-reliability-evidence-melconian-7cpme>
- Qwen Team. (2024, September). *Qwen2.5: A party of foundation models*. <https://qwenlm.github.io/blog/qwen2.5/>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. <https://arxiv.org/abs/1908.10084>
- Rotou, O., & Rupp, A. A. (2020). Evaluations of automated scoring systems in practice (ETS Research Report No. RR-20-10). *ETS Research Report Series*. <https://doi.org/10.1002/ets2.12293>
- Sentence Transformers. (n.d.-a). *all-mpnet-base-v2*. Hugging Face. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>
- Sentence Transformers. (n.d.-b). *all-roberta-large-v1*. Hugging Face. <https://huggingface.co/sentence-transformers/all-roberta-large-v1>
- Tunstall, L., Reimers, N., Jo, U. E., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts: SetFit. *arXiv*. <https://arxiv.org/abs/2209.11055>
- Xue, M., Xiao, X., Liu, Y., & Wilson, M. (2026). On the consistency of automatic scoring with large language models. *Educational and Psychological Measurement*. <https://doi.org/10.1177/00131644261418138>

Appendix A Model Specifications

Automated scoring combined generative and embedding-based approaches. A large language model (Qwen2.5-14B; Qwen Team, 2024) was used for direct scoring, while sentence-transformer models (all-mpnet-base-v2 and all-roberta-large-v1; Sentence Transformers, n.d.-a, n.d.-b) were fine-tuned using SetFit (Tunstall et al., 2022) to learn task-specific scoring functions from labelled responses, building on the Sentence-BERT framework (Reimers & Gurevych, 2019). Model configurations, training sample sizes, and prompt conditions are summarised in Table A1.

Setting	Model	Approach Type	Training Samples	Training Sample Set	Prompt / Input Configuration
1	all-mpnet-base-v2	Sentence Transformer	100	Sample 1	Encoded response only
2	all-mpnet-base-v2	Sentence Transformer	125	Sample 2	Encoded response only
3	all-roberta-large-v1	Sentence Transformer	125	Sample 2	Encoded response only
4	Qwen2.5-14B	LLM	5	Sample 3	Question stem + scoring criteria
5	Qwen2.5-14B	LLM	5	Sample 3	Question stem + scoring criteria + stimulus

Table A1: Model configurations included in the ensemble automated scoring approach, showing variation in model type, training data, and prompt design across components.¹

¹ Note. Sentence transformer models were fine-tuned using SetFit. “Training Samples” refers to number of samples per score category. “Training Sample Set” denotes distinct subsets of audited responses, with some overlap across configurations as described in the Methods section.

Appendix B Detailed Marker Results

Marker Label	Number of Items	Total Scripts	Mean Human Accuracy	Mean AutoScore Accuracy	Difference
M01	30	296	88.8%	95.0%	6.3%
M02	51	457	92.2%	95.4%	3.3%
M03	55	477	92.6%	95.3%	2.7%
M04	40	387	92.7%	95.3%	2.6%
M05	54	472	93.2%	95.6%	2.4%
M06	55	477	93.1%	95.3%	2.2%
M07	55	477	94.0%	95.3%	1.3%
M08	55	477	94.1%	95.3%	1.2%
M09	31	301	94.5%	95.2%	0.7%
M10	23	229	93.9%	94.4%	0.5%
M11	49	432	94.2%	94.7%	0.5%
M12	23	242	94.8%	95.2%	0.4%
M13	21	195	95.4%	95.1%	-0.3%
M14	43	418	95.4%	95.6%	0.3%
M15	36	348	94.8%	95.0%	0.2%
M16	34	338	95.5%	95.3%	-0.2%
M17	50	445	95.5%	95.3%	-0.1%
M18	31	301	95.2%	95.2%	0.0%
M19	25	233	98.2%	95.6%	-2.6%
M20	26	181	95.8%	93.9%	-1.9%
M21	26	180	96.1%	95.2%	-0.9%
M22	29	266	96.7%	96.1%	-0.6%

Table B1: Human marker control script scoring accuracy details with accompanying AutoScore comparison.