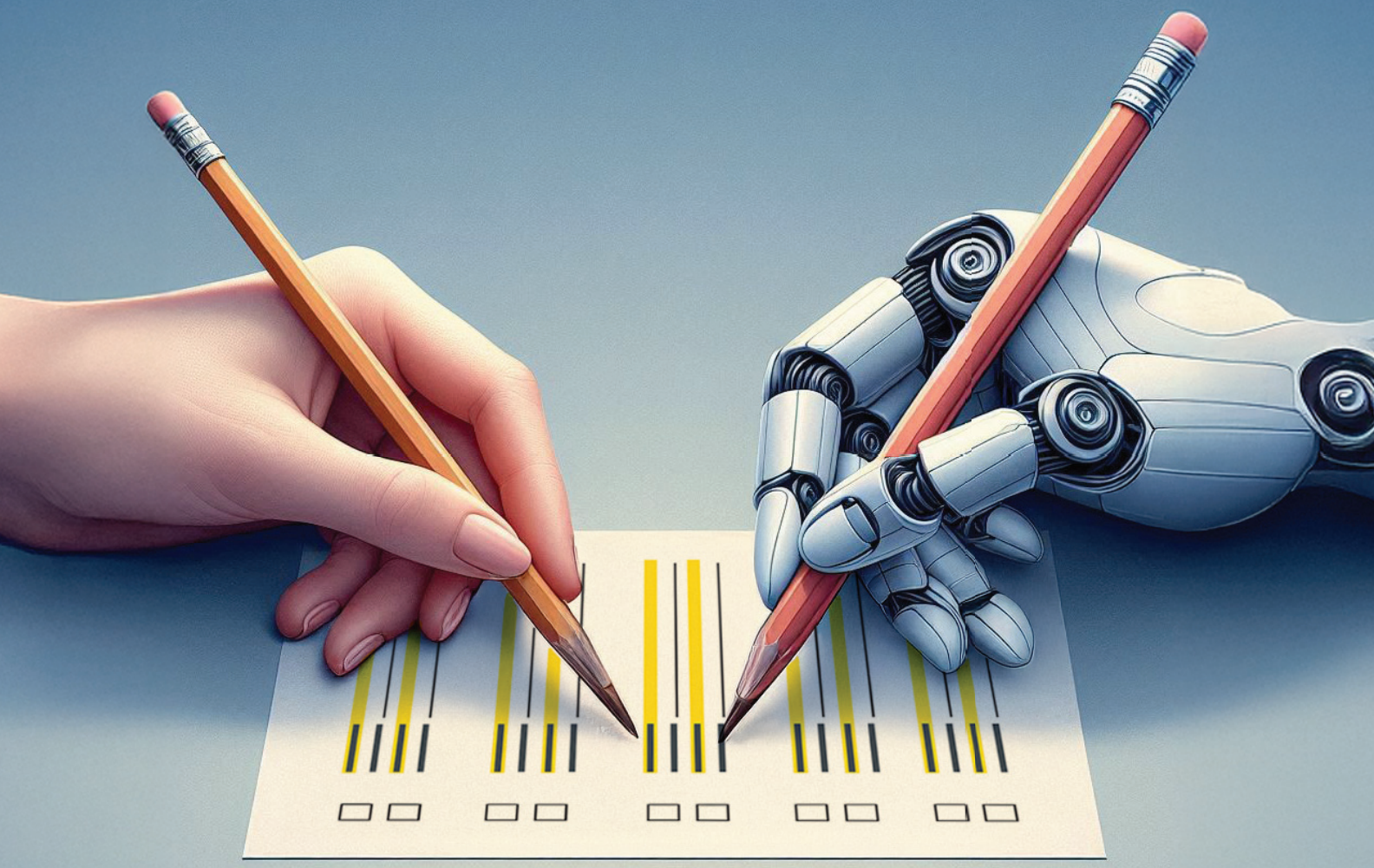
E-SCORING
of Constructed Responses



RICHARD JONES

August 18, 2025

Introduction

Educational large-scale student assessments rely on two general item types: selected-response (e.g., multiple-choice) and constructed-response items, which encompass both short-answer and long-answer formats. In the case of computer-based assessments, selected responses can be automatically machine-scored. Constructed-response formats require test takers to write their answers in both textual and numerical forms. Human scoring of constructed responses involves the potential for subjectivity and the risk of inconsistent, unreliable and invalid results (although it must be recognized that implementing strict scoring procedures can mitigate these issues). Hand scoring also requires significant resources. According to Huseyn (2024), "These two factors, reducing subjectivity and improving operational efficiency, are key reasons behind the growing interest in using technology for automated writing evaluation."[1] This article provides a brief synopsis of the current state of e-scoring for constructed responses and the extent to which it is being considered as part of large-scale student assessment modernization plans in Canada.

Current State of E-Scoring

The automated scoring of constructed responses has undergone significant evolution over the past three decades, and recent advancements in generative AI and natural language processing (NLP) have contributed to the improvement of technology for scoring purposes. Huseyn explains that "…AI-driven developments in the education assessment industry, specifically in scoring and feedback reporting, such as Microsoft Copilot or any OpenAI GPT-based tools, are being considered as integrated assistants to meet the need of marking. Alternatively, some assessment organizations may prefer to go beyond using integrated assistants and instead opt for specialized automated writing evaluation solutions available in the K-12 market designed specifically for this purpose, with fully controlled evaluation algorithms." Huseyn's research paper provides a comparative analysis (a framework) to support decision makers in choosing between GPT-based and specialized scoring systems.

Numerous studies have demonstrated the increasing potential of AI in scoring open-ended responses. For example, a study by Okubo et al. (2023) examined the use of AI to score Programme for International Student Assessment (PISA) constructed-response items.[2] A vast, historical collection of texts (reading and science literacy items) was used as training data to develop AI scoring models. The trained models were then used to score student responses, and the results were compared with those from human-scored responses. The researchers reported that, "The score distributions estimated based on the AI-scored data and the human-scored data are highly consistent with each other; furthermore, even item-level psychometric properties of the majority of items showed high levels of agreement…, and this new AI scoring methodology reached a practical level of quality, even in the context of an international large-scale assessment."

A study by Atsushi and Eguchi (2023) explored the use of ChatGPT for automated essay scoring.[3] In the study, the researchers used 12,100 English essays written by individuals who took the Test of English as a Foreign Language (TOEFL) test in 2006 and 2007 and who represented 11 distinct native languages. Ultimately, the study involved 1,100 essays per language. Specifically, the study used OpenAI's text-davinci-003 model. The results showed that automated essay scoring (AES) "…using GPT has a certain level of accuracy and reliability, and could provide valuable support for human evaluations. Furthermore, the analysis revealed that utilizing linguistic features could enhance the accuracy of scoring. These findings suggest that AI language models, such as ChatGPT, can be effectively utilized as AES tools, potentially revolutionizing methods of writing evaluation and feedback in both research and practice."

[1] Huseyn, V. (2024, October 31). Evolution of e-marking: automated writing evaluation. *Vretta Buzz*. Retrieved July 17, 2025, from: https://www.vretta.com/buzz/automated-writing-evaluation/.

[2] Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, C.S., & Bastianic, T. (2023). AI scoring for international large-scale assessments using a deep learning model and multilingual data. *OECD Education Working Papers, No. 287*, OECD Publishing. Retrieved July 17, 2025, from: https://doi.org/10.1787/9918e1fb-en.

[3] Atsuchi, M., & Eguchi, M. (2023, August). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2 (2). Retrieved July 29, 2025, from: https://www.sciencedirect.com/science/article/pii/S2772766123000101.

An example of AI's application in scoring comes from the United Kingdom (UK). As reported by Pinkstone (2025),[4] the AQA examination board is collaborating with King's College London to develop AI technology designed to support exam paper scorers. In this project, AI is not meant to replace scorers. Instead, the purposes and potential applications are to:

- "…reduce errors, make mark schemes fairer, and give quicker feedback to students."
- "…check the marks given by a human and detect any scores which seem erroneously low or high…"
- "…check the quality of answers from students using language analysis machine learning and scrutinizing the relevance, factuality, coherence and logical reasoning of an answer."
- "…refine the mark scheme if there is a flaw in how marks are given to ensure fairer marking and also give AI-generated explanations to students as to why they did or did not get a question correct."

The project leaders provide assurances that the virtual scoring assistant will be developed in collaboration with students, teachers, and subject experts. The system will be thoroughly tested before being launched, and the constructed responses will always be reviewed by human scorers as well.

In the United States (US), several jurisdictions, including Utah, Ohio, Massachusetts and Texas, utilize AI for various assessment scoring purposes, and many others are exploring AI's potential for constructed response scoring. Massachusetts, for example, uses AI scoring for specific tests. According to the Massachusetts Department of Elementary and Secondary Education (dese)[5], human scorers are used for all constructed-response questions in Civics, English Language Arts (ELA), Mathematics, and Science and Technology/ Engineering. A combination of human scorers and AI scoring is used for ELA essays. For ELA constructed responses (Grades 3 and 4) and essays in Grades 3-8, automated computer scoring provides an initial score. Trained human scorers provide a read-behind score on 10% of all responses. For ELA essays in Grade 10, all essays are scored twice, once by AI and once by a human scorer.

Beginning in 2024, students taking the State of Texas Assessments of Academic Readiness (STAAR) tests had their written answers for reading, writing, science and social studies scored automatically by computers.[6] The decision to transition to computer scoring was made following a redesign of the tests, which involved reducing the number of multiple-choice items and significantly increasing the number of constructed-response questions/tasks. In developing the scoring system, the Texas Education Agency (TEA) utilized a field sample of 3,000 student responses that underwent two rounds of human scoring. From this sample of scored responses, the automated scoring engine learned the characteristics of the responses and was programmed to assign scores identical to those a human would have given. As students completed their tests, the scoring system first graded all the constructed responses. Approximately 25% of them were rescored by human scorers. In instances where the computer system had low confidence in its assigned scores, they were reassigned to human scorers for review. Similarly, when the system encountered a type of response that its programming did not recognize (e.g., use of many slang words/expressions, words in languages other than English), it was reassigned to a human. Additionally, a random sample of responses was automatically forwarded to humans for verification of the automated system's accuracy.

[4] Pinkstone, J. (2025, July 12). AI to help mark student exams. *The Telegraph*. Retrieved July 24, 2025, from: https://www.yahoo.com/news/ai-help-mark-student-exams-160000080.html?guce_referrer=aHR0cHM6Ly9uZXdzLmdvb2dsZS5jb20v&guce_referrer_sig=AQAAALF8DlkmGCyU7iF6U5z3m6Z5we_BaNowjGPWZctr0H8nZw9EVj4Z-8DJL0EfEEEUP6dL48YU2UwbI1rAkBEKyfZBArsvKLlbZxa6mvjcXwQ69ut5X0EkoZQP4RlbrA6OAm3xHT74dHqy5SCMHdN5vsxuk6Jkkv8-xfbyjwonieLS&utm_source=TestCommunityNetwork&_guc_consent_skip=17533697014.

[5] Massachusetts Department of Elementary and Secondary Education. (2025, July 18). Massachusetts Comprehensive Assessment System: *Scoring Student Answers to Constructed-Response Questions and Essays*. Retrieved July 28, 2025, from: https://www.doe.mass.edu/mcas/student/2024/scoring.html#:~:text=two%20separate%20scorers.-,Automated%20(Computer)%20Scoring,metrics%20and%20requirements%20are%20met.

[6] Peters, K. (2024, April 9). Texas will use computers to grade written answers on this year's STAAR tests. The Texas Tribune. Retrieved July 28, 2025, from: https://www.texastribune.org/2024/04/09/staar-artificial-intelligence-computer-grading-texas/.

The Canadian Context

In Canada, the majority of jurisdictions (provinces and territories) have transitioned their large-scale student assessment programs from paper-based to computer-based formats, with Vretta serving as the technology partner supporting this modernization. These jurisdictions have implemented digital platforms that include online human scoring and integrated workflows designed to support large-scale assessment delivery.

Building on this foundation, three provinces have begun benchmarking, trialling, and implementing AI scoring within their assessment programs. These jurisdictions, already leading in modernized assessment practices, are now focused on leveraging AI to support scorers in their work and enhance the quality and consistency of scoring through intelligent validation processes. The use of AI in these contexts is not to replace human judgment, but rather to increase operational efficiency, reduce scoring discrepancies, and support continuous improvement in the reliability of student assessment results.

This next wave of innovation highlights Canada's commitment to responsible and progressive use of technology in education, with a strong emphasis on quality, fairness and validity in the scoring process. I believe that the momentum toward integrating AI in scoring is expected to grow as other jurisdictions observe these early implementations and evaluate the benefits within their own contexts.

Conclusion

Research has revealed the great potential and benefits of AI scoring, and increasingly, jurisdictions and testing organizations are either implementing or exploring the use of technology for the scoring of open-response questions/tasks. A future article will report on the progress toward the implementation of automated constructed-response scoring in Canada.

About the Author

Dr. Jones has extensive experience in the fields of large-scale educational assessment and program evaluation. He has worked in the assessment and evaluation field for nearly 40 years. Before founding RMJ Assessment, he held senior leadership positions with the Education Quality and Accountability Office (EQAO) in Ontario, as well as the Saskatchewan and British Columbia Ministries of Education. In these roles, he was responsible for initiatives related to student, program and curriculum evaluation; education quality indicators; school and school board improvement planning; school accreditation; and provincial, national and international testing.

Dr. Jones began his career as an educator at the elementary, secondary and post-secondary levels. Subsequently, he was a researcher and senior manager for a multi-national corporation delivering consulting services in the Middle East.

Feel free to reach out to Richard "Rick" at richard.jones@rmjassessment.com (or on LinkedIn) to inquire about best practices in large-scale assessment and program evaluation.