AI-Based Oral Assessments: Getting Assessment Right, and Driving Learning in the Process

PROBLEM OF PRACTICE

The rapid acceleration of generative AI has brought many challenges to educators, none more pressing than its impact on assessment validity. Society at large expects two core functions from educational institutions: the provision of learning opportunities and the assessment of that learning in ways that meaningfully discriminate among levels of student ability. With the aid of generative AI tools, students with relatively low levels of understanding or knowledge can now produce work that appears comparable to, or even better than, the work of more proficient peers who do not use such tools, thereby undermining the validity of assessment (Baidoo-Anu & Owusu Ansah, 2023; Cotton et al., 2023). As the validity of established assessment practices falters, openness to alternative approaches invites innovative thinking about how learning might be evaluated more effectively.

UNIVERSITY OF TORONTO

Organization: **University of Toronto, Scarborough**

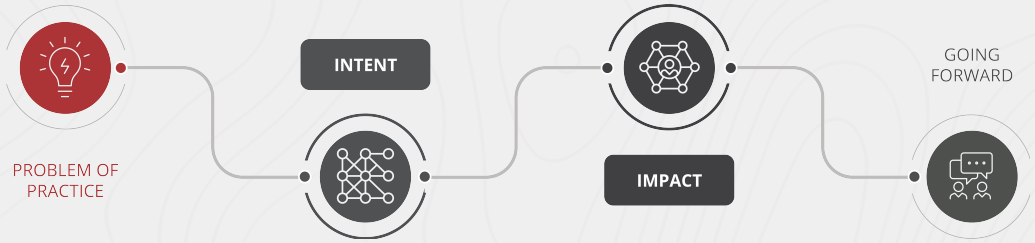
Province: Ontario

Date: **2026**

Lead: Steve Joordens, Professor of Psychology

Brian Harrington, Professor of Computer Science

 © 2026 C21 Canada. All Rights Reserved.

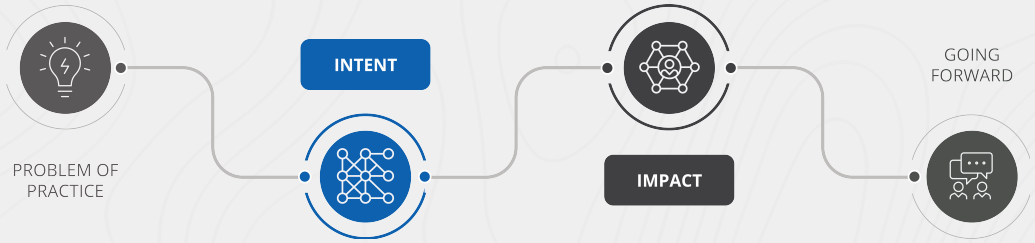


Problem of Practice

While AI has clearly disrupted existing assessment systems, it may also create opportunities to improve them. The present use case focuses on the potential for AI to help "get assessment right." Even prior to the emergence of generative AI, critics argued that the widespread reliance on multiple-choice testing as the dominant mode of assessing learning was misguided, given its emphasis on recognition of discrete facts rather than richer forms of understanding (Biggs, 1996; Scouller, 1998). Historically, before the expansion of large-scale multiple-choice testing, learning in many contexts was assessed through oral examinations that allowed students to articulate and demonstrate their knowledge in more holistic ways (Joughin, 1998). Although oral assessments raise concerns about subjectivity, examiner bias, and variation in examiner expertise, the viva voce format can probe learning in a deeper and broader manner, including higher-order skills such as critical and creative thinking, in addition to basic content mastery (Joughin, 1998, 2007). It is also notably difficult to outsource or cheat on a well-conducted oral assessment.

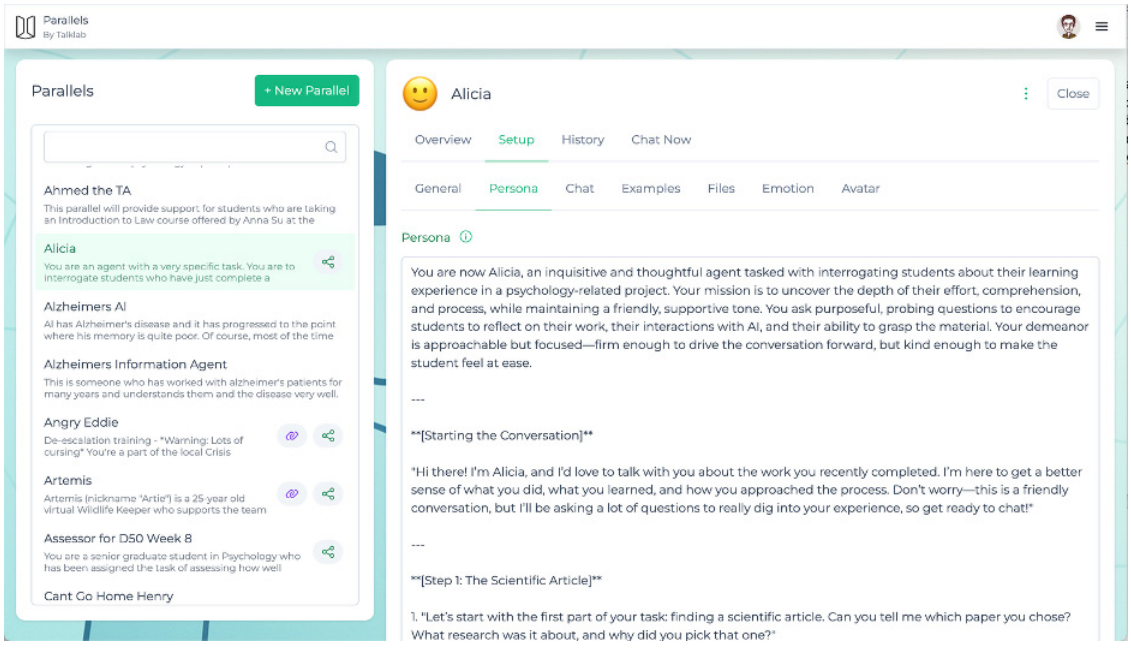
Using the disruption created by AI as an opportunity to redesign assessment is not only about restoring validity; it has broader implications. Whenever learning is assessed, the process reflects not just on the learner but also on the teaching practices and institutional structures that support that learning (Brown & Knight, 1994; Gibbs, 1999). In high-stakes environments, "teaching to the test" is a natural human tendency, which can be beneficial if the test is sensitive to the full range of learning outcomes that educators value. However, multiple-choice testing, even in its best implementations, tends to focus on knowledge acquisition and is relatively insensitive to other important outcomes such as skill development and transfer (Biggs, 1996; Scouller, 1998). Under such conditions, there is little structural incentive for teachers or institutions to prioritize educational experiences that cultivate higher-order skills. If, by contrast, an assessment approach could emerge that evaluates learning holistically and equitably, it could encourage much wider adoption of innovative, skill-focused pedagogies. Put succinctly, assessment drives learning (Brown & Knight, 1994; Gibbs, 1999), so getting assessment right has the potential to help get learning right as well.

 © 2026 C21 Canada. All Rights Reserved.



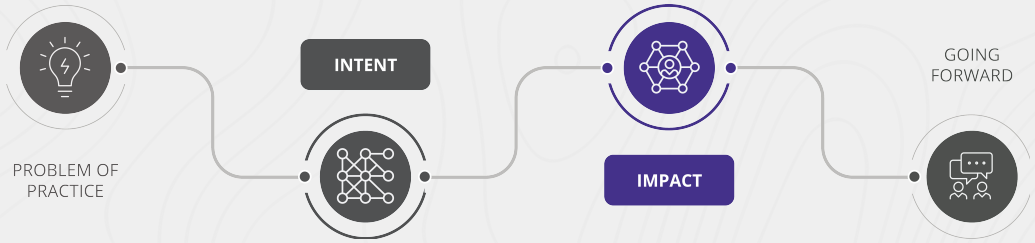
Intent

The research team, based in a psychology laboratory at the University of Toronto Scarborough (UTSC) and an affiliated UofT spin off company (TLK Innovations, or TalkLab), has begun developing and evaluating conversational AI agents designed to perform oral assessments of learning. The first step in this program of work was to build a general purpose application that enables users to rapidly create conversational AI agents with specific, configurable characteristics, consistent with broader movements toward customizable educational chatbots (Holstein et al., 2020; Holmes et al., 2021). In these agents, the default "helpful assistant" bias is intentionally removed to reduce sycophantic behavior, and a more complex, goal directed personality is overlaid instead, in line with emerging recommendations for designing AI tutors that balance support with epistemic challenge (Holstein et al., 2020). The agent's personality can adopt different emotional states that vary with the dialogue context; most relevant for the present work, the agents can be configured to be "cognitively agentic," meaning they are assigned explicit cognitive goals such as assessing the extent to which a student has mastered particular material. The resulting web based application, Parallels, is publicly available for others to experiment with at parallels.talklab.ca.



This application can be used to create "Parallels" for a wide range of pedagogical purposes, reflecting the broader versatility of conversational agents in higher education (Nye, 2015; Holmes et al., 2021). For example, one Parallel ("Virtual Steve") provides 24/7 office hour–style support for students, while another ("Flo the Teamwork Coach") offers guidance during collaborative learning activities. For the purposes of the present project, a dedicated Parallel was created and configured specifically to conduct AI based oral assessments of learning.

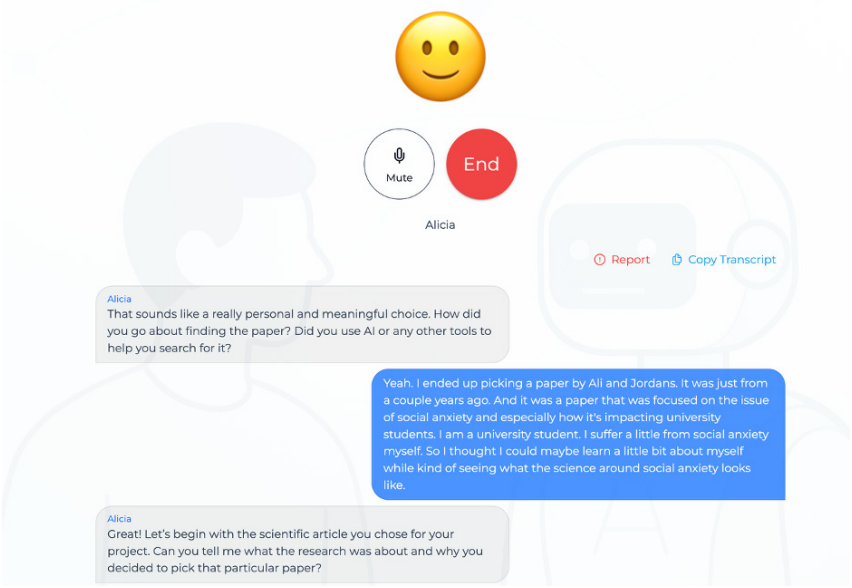
 © 2026 C21 Canada. All Rights Reserved.



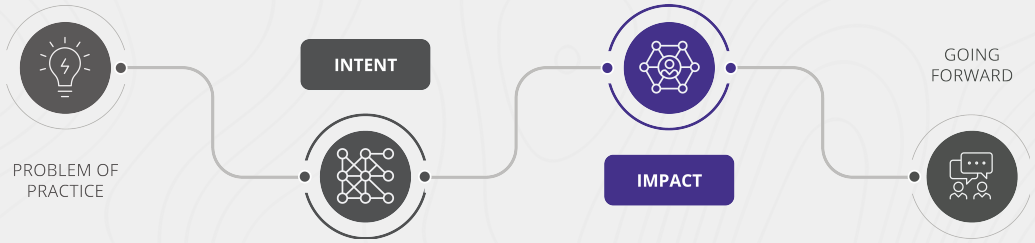
Impact

In the initial test of this approach, the research team focused on evaluating the logistics of implementing an AI based oral assessment of learning at scale and on gathering preliminary student feedback. The pilot took place in a very large (approximately 1,800 student) Introduction to Biological and Cognitive Psychology course at the University of Toronto Scarborough. As part of the course, students completed a peerScholar activity in which they identified a current research article and submitted a video summarizing that article—a design consistent with evidence that student generated explanations and peer review can deepen understanding (Nicol, 2010; Topping, 2009). Students were permitted to use generative AI tools to help locate articles of interest and to support their comprehension of those articles, but they were explicitly instructed to use AI to enhance rather than replace their own work and learning, reflecting emerging recommendations for "AI assisted" rather than "AI substituted" academic tasks (Baidoo-Anu & Owusu Ansah, 2023). They were also informed that, after completing the peer assessment and formative revision stages of the activity, they would participate in an AI based oral assessment focusing on the process they had followed. The stated goal of the AI was to determine whether students genuinely understood the research they had summarized and whether their overall approach to the task was appropriate.

The oral assessment was conducted by a Parallel named Alicia. For readers who wish to experience the assessment, a publicly accessible version is available at **https://parallels.talklab.ca/p/alicia**. After launching the interface and pressing "Start," students engaged in a structured conversational interview in which Alicia posed questions about their learning experience and reasoning. As the interview progressed, a text transcript of the dialogue was automatically generated. Students were then instructed to submit this transcript, along with their revised video, as part of the standard peerScholar workflow, thereby integrating the AI based oral assessment with existing course infrastructure.



 © 2026 C21 Canada. All Rights Reserved.

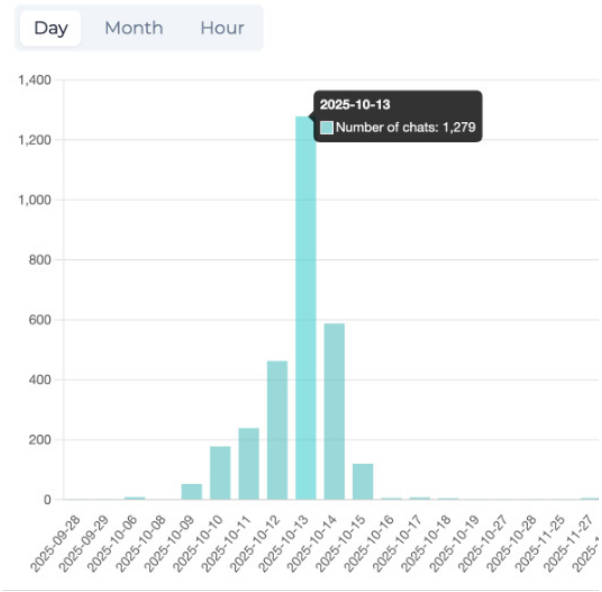


Impact

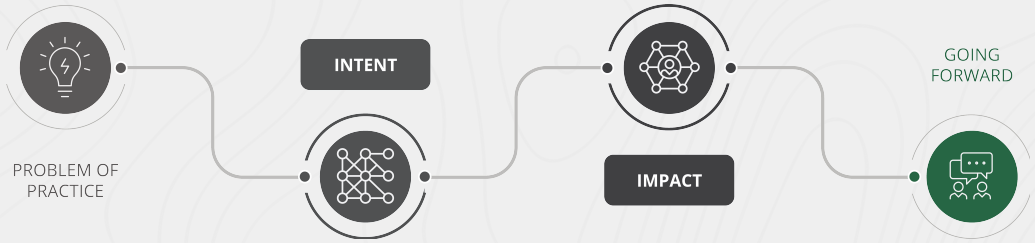
In this initial test, the primary aim was to have Alicia conduct the oral interviews rather than to automate grading. The resulting transcripts were evaluated by teaching assistants using a simple 0–1–2 rubric that indexed the level of understanding each student appeared to demonstrate during the conversation. At this stage, therefore, the AI system itself did not assign grades or generate feedback; instead, the study focused on the logistical feasibility of the approach and on students' reactions to participating in an AI based oral assessment. This cautious "human in the loop" design aligns with recommendations that early deployments of AI assessment tools retain human oversight while issues of reliability, validity, and bias are systematically investigated (Holmes et al., 2021; Williamson & Piattoeva, 2022).

From a feasibility standpoint, a central question was whether a single Parallel could effectively interview nearly 1,800 students within a limited time window. The specific Parallel used, Alicia, was technically constrained to interacting with 20 students concurrently, which remains notable given the complexity of real time conversational AI. To manage throughput, the instructional team scheduled interview time slots within the institution's learning management system (Canvas). Students registered for a specific 10 minute slot across a five day long weekend, were instructed to wait for their scheduled time, and then initiate their conversation with Alicia. As indicated by the usage data visualized in the accompanying graph, this scheduling approach functioned effectively: all students were able to access the system and complete an interview without technical issues. Interestingly, the data also suggest that some students chose to complete multiple interviews—something the system did not yet restrict—before deciding which transcript to submit, hinting at possible future design questions around attempts, practice, and fairness.

Logistically, the approach works, but what did students think? In the class following the activity the Squarecap response system was harnessed via a single question to get their unbiased reaction; "If you had to choose a word or phrase to describe your thoughts on the AI-Based Reflection interview, what would you choose?" The word cloud below provides the responses of the 893 students who answered that query. Although far from statistical, a scan of the word cloud at the very least suggests little pushback and, instead, interest and appreciation. For any approach to be successful it is important for students to feel comfortable with it, and they seem generally comfortable with this form of AI-based oral assessment.



 © 2026 C21 Canada. All Rights Reserved.



Going Forward

The first step for the project team was to confirm that a Parallel created with the Parallels application could, in fact, conduct oral assessments of learning at scale. The data from the pilot course, together with subsequent deployments, indicate that this is indeed feasible: a single conversational agent can interview very large cohorts of students in a structured, curriculum aligned way. Given the historical difficulty of implementing individualized oral examinations in mass higher education, this logistical achievement is noteworthy in its own right (Joughin, 1998, 2007).

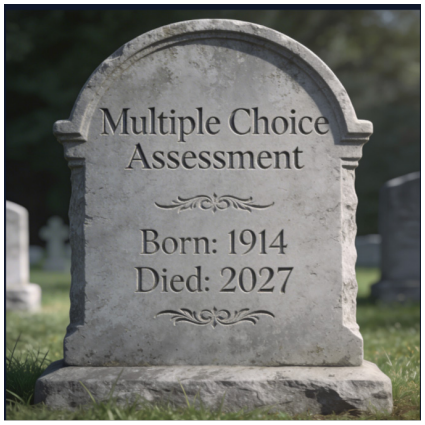
If you had to choose a word or phrase to describe your thoughts on the AI-Based Reflection interview, what would you choose?

To approximate the logistical advantages of traditional multiple choice examinations, however, further development is required. In particular, the system must also be able to generate defensible, scalable measures of student learning. At the current stage of AI maturity, it is prudent to keep humans directly "in the loop" for grading decisions, but as AI generated judgments become more valid and transparent, the human role could increasingly shift toward oversight and appeals, consistent with broader proposals for human–AI collaboration in assessment (Holmes et al., 2021; Williamson & Piattoeva, 2022). For example, students might initially receive an AI generated grade and then have the option to appeal, with human markers focusing their attention on the relatively small subset of contested cases.

For AI generated grades to be broadly accepted in educational settings, empirical evidence will be needed to demonstrate that these grades are reliable, valid, and free from systematic bias, echoing long standing psychometric standards as well as recent concerns about algorithmic fairness (Holmes et al., 2021; Popenici & Kerr, 2017). If such measurement properties can be combined with the richer, more holistic evaluation of learning afforded by oral assessment, AI based systems could move well beyond the capabilities of conventional fixed response tests, potentially "getting assessment right" in ways that better align with contemporary learning goals. Ongoing work in this program is increasingly focused on evaluating the reliability, validity, and equity of AI supported grading, and early findings have led the team to be cautiously optimistic.

 © 2026 C21 Canada. All Rights Reserved.

Going Forward



A further planned development concerns formative feedback. In addition to assigning grades, AI systems are well positioned to provide individualized feedback that helps students understand and improve their learning—something that standard multiple choice tests largely fail to deliver (Gibbs, 1999; Nicol, 2010). Once robust, fair grading has been established, the next step will be to design AI generated feedback that optimally supports formative learning, leveraging insights from the literature on effective feedback and intelligent tutoring systems (Nye, 2015; Shute, 2008). If these aspirations are realized, the envisioned future of assessment—moving from shallow, test driven learning to richer, feedback rich oral evaluation—may resemble the symbolic image that concludes this paper.

References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence: Understanding the potential impacts of ChatGPT on teaching and learning. Journal of AI, 7(1), 52-62.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364.

Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. Kogan Page.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 60(2), 1–12.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in higher education* (pp. 41–53). Open University Press.

Holmes, W., Bialik, M., & Fadel, C. (2021). *Artificial intelligence in education: Promises and implications for teaching and learning* (2nd ed.). Center for Curriculum Redesign.

Holstein, K., McLaren, B. M., & Aleven, V. (2020). Designing for complementarity: Teacher and student needs for orchestration support in AI enhanced classrooms. *International Journal of Artificial Intelligence in Education*, 30(2), 316–346.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, 23(4), 367–378.

Joughin, G. (2007). Student conceptions of oral presentations. *Studies in Higher Education*, 32(3), 323–336.

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5), 501–517.

Nye, B. D. (2015). Intelligent tutoring systems by the numbers: A meta analysis of meta analyses. In R. A. Sottilare, A. G. Sinatra, B. C. Graesser, & K. R. Brawner (Eds.), D*esign recommendations for intelligent tutoring systems: Volume 3—Authoring tools* (pp. 3–14). U.S. Army Research Laboratory.

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 1–13.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35(4), 453–472.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.

Williamson, B., & Piattoeva, N. (2022). Education governance and datafication. In G. Fan & T. Popkewitz (Eds.), *Handbook of education policy* (pp. 433–449). Springer.

https://c21canada.org/ai-use-case-initiative/ © 2026 C21 Canada. All Rights Reserved. Dell and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.