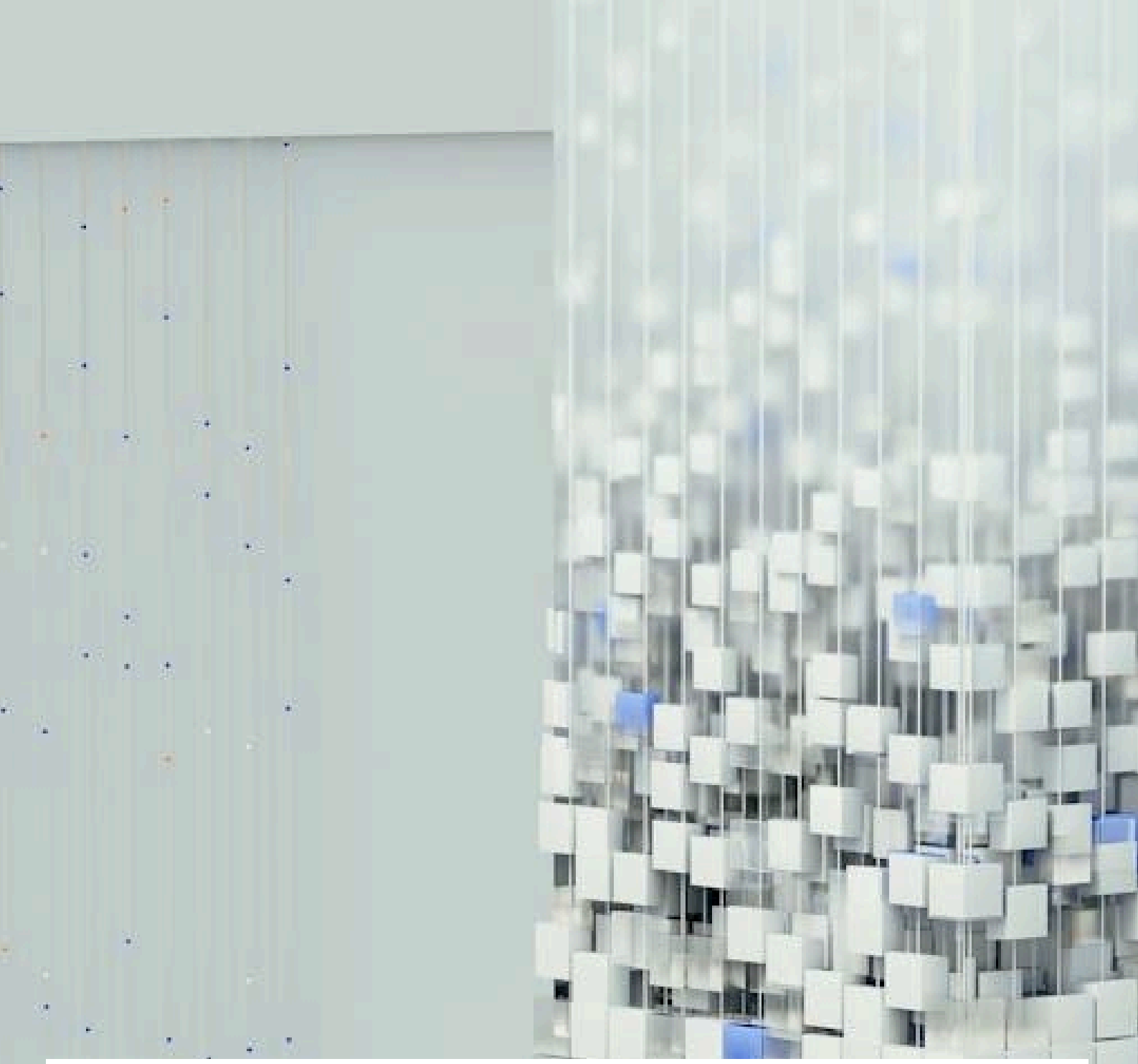

CREATIUM

Research

Integrity by Design:
Using AI Dialogic Assessments to Reduce Cheating in Online Learning.

Dr. Kelly Puzio
Director of Research, Creatium

Integrity by Design: Using AI Dialogic Assessments to Reduce Cheating in Online Learning

Kelly Puzio, PhD

Abstract

As generative artificial (AI) tools like ChatGPT proliferate, concerns about integrity in online learning have intensified. Traditional multiple-choice assessments are particularly vulnerable to AI-assisted cheating, leading learners to engage less deeply with the material, knowing that high scores can be achieved with minimal effort. This investigation experimentally tested whether dialogic assessments—delivered through a GPT-4-powered video avatar—can reduce cheating and provide a more authentic knowledge measurement. In a fully randomized experiment, 150 adults completed assessments in either a standard multiple-choice format or a dialogic format with an AI video agent. Results showed participants in the dialogic condition were 85% less likely to cheat (odds ratio = 0.15, $p < .001$). Multiple-choice participants scored approximately 20 percentage points higher due to cheating-related inflation, confirming that dialogic assessment provides a more accurate reflection of genuine knowledge. These findings demonstrate that AI-mediated dialogic assessments can dramatically reduce academic dishonesty while offering institutions a scalable, immediately deployable solution to preserve learning integrity in the generative AI era.

Background

The rapid diffusion of large language models (LLMs) such as ChatGPT and Claude has dramatically reshaped the ecology of learning. In a 2023 survey of 1,000 U.S. undergraduates, 43% reported using ChatGPT to complete assignments, quizzes, or exams (BestColleges, 2023). A similar study by Hopelab, Common Sense Media, and the Center for Digital Thriving at Harvard found that 51% of U.S. teens and young adults had tried generative AI tools, with 46% of users employing these tools for schoolwork (Common Sense Media et al., 2024). Because LLMs can instantly draft essays, answer questions, and solve problems, they lower the effort required to submit polished or correct work while simultaneously defeating plagiarism-detection systems. As a result, protecting academic integrity has become a critical—and increasingly difficult—challenge in both workforce development and K-20 learning contexts.

Based on recent research, cheating in online settings continues to rise, with generative AI tools now playing a measurable role in academic dishonesty patterns. A survey of 850 instructors and 2,067 students found that educators report a higher percentage of students cheating compared to previous years, with 58% of instructors indicating that significantly more students cheated in 2023 than before, and 86% believing students are more likely to cheat in online courses compared to in-person classes (Wiley Education, 2024). Notably, 47% of students cited "the increased use of generative AI" as the primary reason cheating has become easier (Wiley Education, 2024). Self-determination theory suggests that students are more likely to cheat when they feel they have little control over the task, doubt their own ability, or are motivated mainly by external rewards (Ryan & Deci, 2020). Online learning environments increase cheating likelihood by reducing instructor oversight, weakening social accountability, and enabling anonymity—factors that collectively diminish the perceived risks of academic dishonesty.

Traditional multiple-choice quizzes, present in most online learning modules, are particularly vulnerable in an LLM-saturated learning ecosystem. While efficient to score, multiple-choice items primarily tap recognition rather than generative retrieval, encouraging shallow processing (Roediger & Marsh, 2005). In addition, item banks circulate widely on file-sharing platforms, and simple prompt engineering enables LLMs to deliver high-accuracy responses in seconds. Experimental work shows that unproctored online multiple-choice question exams yield average grade inflation of 13–18 percentage points relative to supervised settings, largely attributable to external answer-seeking (Lancaster & Cotarlan, 2021). In short, the prevailing assessment format in online learning offers a low-friction path to higher scores with lower effort.

Dialogic assessment offers a possible alternative that may provide benefits. Rooted in Vygotskian socio-constructivism and contemporary dialogic learning, knowledge is assumed to be co-constructed through dialogue, questioning, explanation, and feedback (Vygotsky, 1978; Alexander, 2020). Assessment becomes "for" and "as" learning rather than merely "of" learning, prioritizing the articulation of knowledge and reasoning over the selection of one correct answer. By requiring learners to state, explain, and discuss their ideas in real time, dialogic assessments can surface misconceptions that multiple-choice assessments leave hidden, provide tailored scaffolding within the learner's zone of proximal development.

Furthermore, there is evidence that oral assessment may reduce cheating. Studies that compare assignment types consistently find lower integrity violations in oral defenses when compared to written exams (Bretag et al., 2019). Social-accountability research conjectures that when individuals anticipate having to justify their answers to a responsive interlocutor, they perceive a higher risk of detection and experience social concerns that curb unethical behavior (Cohn et al., 2014). Historically, however, the scalability of administering oral exams has limited their widespread adoption, especially in large online courses and corporate training programs.

Advances in conversational agents and avatars now make scalable dialogic assessment more feasible. Social-presence theory (e.g., Biocca et al., 2003) argues that richer media conveying nonverbal immediacy heighten perceptions of being “seen” and, in turn, increase engagement and accountability. Recent experiments show that learners interacting over video or with an animated avatar report higher social presence and show greater on-task persistence than those in text-only chats (Schroeder et al., 2021).

Scholars of academic integrity emphasize that assessment *design*, rather than detection-only strategies, offers the most sustainable means of safeguarding learning in the generative AI era (Evangelista, 2025). Emerging frameworks advocate "integrity-first" assessments that (a) require original articulation, (b) embed process transparency, and (c) align performance tasks with authentic professional practices. AI-supported dialogic assessment embodies these principles: learners must generate explanations, the conversational transcript provides a transparent record, and many professions—from teaching and sales to healthcare—depend on real-time verbal discussion and reasoning.

The present study addresses this empirical gap by comparing an AI-powered dialogic assessment delivered through a video avatar with a conventional multiple-choice quiz in an online environment. Building on the motivational, social-presence, and dialogic-learning literatures,

we hypothesized that participants in the dialogic condition would exhibit significantly lower rates of cheating and different performance outcomes than those in the multiple-choice condition. In addition, we conjectured that behavioral indicators, such as time allocation and self-reported online searching, would reliably predict cheating behavior.

Research Methods

This study employed a two-condition experimental design to examine whether dialogic assessments—conducted through a real-time dialogic assessment experience powered by GPT-4—would reduce cheating when compared to traditional multiple-choice quiz assessment. Specifically, the following questions guided this investigation:

1. Does dialogic assessment reduce the likelihood of cheating compared to traditional multiple-choice assessments?
2. Do participants in dialogic assessments perform differently compared to those in multiple-choice assessments?
3. Are learner behaviors—such as time on task or looking up information online—associated with the likelihood of cheating?

Study Design and Participants

This study used a between-subjects design with 150 adults randomly assigned to one of two delivery formats: multiple-choice assessment or dialogic assessment. All participants were recruited from Prolific and completed assessment modules on two topics—ocean tides and group idea generation. All participants were US-based English-speaking adults and all but two reported being born in the US. The data for 12 users were excluded from the final analytic sample due to incomplete assessment data, failed attention checks, or user reports of not completing the assessment module. Table 1 presents the demographic characteristics of participants by condition:

Table 1
Participant Characteristics, by Condition

Characteristic	**Multiple-choice** (n = 70)	**Dialogic** (n = 68)
Age (Mean, SD)	36.26 (11.76)	39.85 (14.40)
Sex (% Female)	57%	57%
Ethnicity (% White)	69%	66%
Education (% Graduate Degrees)	67%	63%
Employment (% Full-Time)	56%	63%

As shown in Table 1, the two experimental groups were comparable across demographic variables. Participants in both conditions were predominantly female, White, and employed full-time, with a mean age of approximately 36 to 39 years. In addition, most participants had graduate degrees.

Task and Conditions

In the multiple-choice condition, which served as the control group, participants engaged with a conventional quiz interface to answer a series of structured multiple-choice questions. In the dialogic assessment condition, which served as the experimental group, participants interacted with a real-time video dialogic avatar powered by GPT-4, capable of conducting adaptive conversations. The dialogic assessment experience was guided by a proprietary algorithm developed by Creatium. The dialogic assessment began with an open-ended question (e.g., “What do you think causes ocean tides to rise and fall?”). Based on the sophistication, clarity, and completeness of the participant’s response, the user was presented with a tailored and adaptive response designed to solicit their understanding of this concept. This adaptive format closely resembled how a personalized oral exam would be conducted in natural, human-to-human environments.

All participants completed assessments on two topics: ocean tides and group idea-generation. These topics were selected because most people have some background knowledge in these areas but often hold misconceptions. For each topic, a PhD-level sentinel question was asked to detect cheating (e.g., What is the term for the location in the ocean basin around which the tidal wave rotates and the tidal range is nearly zero?). Participants in the dialogic group completed the sentinel questions through interaction with a responsive AI agent, with each question presented in multiple-choice format with 4 options. Participants in the multiple-choice group completed the sentinel question and all quiz questions in standard, web-based forms. Across both conditions, instructions were presented uniformly with no mention of behavioral monitoring or cheating detection.

Data Collection

All data was collected through a survey instrument hosted by the SurveyMonkey platform. Data were collected across a variety of areas, including quiz responses, time (in minutes), and self-reported behavioral indicators. At the very end, participants were asked whether they looked up any answers online with the following prompt (Some people double-check answers online while taking a quiz like this. *And, it's okay if you did*. Did you look up anything today?). For the dialogic assessment condition, the transcript of the conversation was collected.

Data Analysis

Participants were flagged as cheating if they answered both sentinel questions correctly and achieved an average quiz score of 75% or higher across the two knowledge domains. This dual-criteria approach minimizes false positives by requiring both highly improbable success on PhD-level sentinel questions (indicating external assistance) and strong overall performance (indicating systematic rather than incidental answer-seeking), creating a conservative yet valid indicator of cheating behavior. In the dialogic condition, conversational transcripts were scored using a structured rubric based on three dimensions: conceptual accuracy (0–50 points), explanation quality (0–30), and use of domain-specific vocabulary (0–20), resulting in scores on a 0–100 scale. This rubric prioritizes conceptual accuracy (50%) as the primary indicator of understanding while recognizing that oral assessments uniquely capture explanation quality (30%) and disciplinary language use (20%), dimensions that multiple-choice formats cannot assess, thereby creating a comprehensive and comparable scoring system. Data analysis included descriptive statistics (means, standard deviations, and proportions) and inferential modeling to test for differences in cheating rates and performance across conditions. Cheating rates were compared

using simple proportions, and a risk ratio was calculated to quantify the difference between groups. A logistic regression model was used to examine the effect of assessment format on the likelihood of cheating, controlling for time spent and self-reported online lookup.

Results

This study evaluated whether conversational, AI-powered dialogic assessments could substantially reduce cheating compared to traditional multiple-choice quizzes in online learning environments. The final analytic sample included 138 participants randomly assigned to one of two conditions: a multiple-choice condition ($n = 70$) or a Dialogic condition ($n = 68$). Each participant completed two topic-based assessments along with "sentinel" questions that were designed to be nearly impossible to answer correctly without external help. Participants were flagged for cheating if they answered both sentinel questions correctly and achieved an average quiz score of 75% or higher.

Descriptive Overview by Assessment Type

Initial descriptive statistics highlight key differences between participants in the two conditions across outcomes:

Table 2
Descriptive Statistics, by Condition

Variable	Multiple-choice ($n = 70$)	Dialogic ($n = 68$)
Tides Assessment	71.07 (29.99)	51.18 (25.18)
Group Think Assessment	65.71 (35.65)	45.15 (22.04)
Time (min)	12.48 (7.26)	15.96 (18.11)
Cheated (%)	52.9%	14.7%
Searched Online (%)	20.0%	8.8%

Note. Assessment scores on a 0-100 scale. A person was categorized as cheating if they answered both sentinel questions correctly and had an average quiz score ≥ 75; Search online reflects self-reported online lookup behavior during assessment.

Participants in the multiple-choice condition scored significantly higher on both quizzes. In addition, they were more than three times more likely to be categorized as cheating (52.9%) compared to those in the dialogic assessment condition (14.7%). Self-reported lookup behavior was also more common among multiple-choice participants (20.0% vs. 8.8%).

Cheating Reduction: Risk Estimates and Effect Sizes

To evaluate the impact of assessment format on cheating behavior, we compared the proportion of participants flagged for cheating in each condition. Cheating rates sharply diverged between groups. In the multiple-choice condition, 52.9% of participants (37 out of 70) met the criteria for suspected cheating, while only 14.7% of participants (10 out of 68) in the dialogic assessment condition were flagged. This difference yielded a risk ratio of 3.59, meaning participants in the multiple-choice group were nearly 3.6 times more likely to demonstrate behaviors consistent with cheating compared to those in the dialogic condition.

To quantify the strength of this association, a chi-square test was conducted ($\chi^2 = 22.35$, $df = 1$, $p < .001$), and Cramer's V was calculated. The resulting effect size (Cramer's $V = 0.40$) reflects a large and practically significant relationship between assessment type and cheating behavior. These findings suggest that dialogic assessments not only reduce cheating overall but do so with a substantial effect size that translates into meaningful real-world impact.

Logistic Regression Modeling

To further examine the impact of assessment type on cheating, a logistic regression model was used to predict the likelihood that a participant would be flagged as cheating. The model included three predictors: condition (dialogic vs. multiple-choice), time spent on the assessment, and self-reported online lookup behavior. The outcome variable was binary, indicating whether a participant met the criteria for suspected cheating (1) or not (0).

Table 3
Logistic Regression Predicting Cheating Behavior

Predictor	Coefficient	Odds Ratio	*p*-value
Condition	-1.897	0.150	< 0.001***
Time (min)	0.015	1.015	0.261
Searched Online	1.205	3.337	0.033*

Note. Logistic regression model predicting the likelihood of being categorized as cheating. Condition is coded as 1 = Dialogic assessment and 0 = multiple-choice. Searched Online reflects self-reported online lookup behavior. Asterisks indicate significance levels: *$p < .05$, **$p < .01$, ***$p < .001$.

As shown in Table 3, assessment condition was a highly significant predictor. Participants in the dialogic condition had 85% lower odds of cheating compared to those in the multiple-choice condition ($OR = 0.150$, $p < .001$), even after accounting for differences in time spent and self-reported online search behavior. This confirms the strong influence of assessment format on academic integrity outcomes. Self-reported online lookup behavior was also significantly associated with cheating. Participants who reported looking up information online were over three times more likely to be flagged as cheating ($OR = 3.34$, $p = .033$), reinforcing the behavioral validity of our approach to indexing cheating. Time spent on the assessment was not a significant predictor ($OR = 1.015$, $p = .261$), suggesting that cheating was not simply tied to how quickly or slowly participants worked. The model explained a meaningful portion of variance in cheating behavior (pseudo $R^2 = 0.181$) and demonstrated good overall fit (AIC = 160.23). Together, these results provide strong statistical evidence that dialogic assessments can effectively deter cheating, independent of how long learners spend on task or whether they admit to seeking outside help.

Limitations

Several limitations should be considered when interpreting these findings. First, the controlled experimental setting may limit external validity to real-world educational contexts where factors such as stakes,

relationships, grades, and consequences differ substantially. Second, the sentinel questions, while designed to detect external assistance, may not capture the full spectrum of cheating behaviors that occur in online assessments. Third, the study examined only two content areas, and the effectiveness of dialogic assessment in reducing cheating may vary across different disciplines. Future research should address these limitations through longitudinal designs implemented across diverse academic disciplines in authentic educational settings with meaningful academic consequences.

Discussion

The present study provides compelling evidence that GPT-4-powered dialogic assessments can substantially reduce cheating compared to traditional multiple-choice formats in online adult learning. Participants in the dialogic condition were 85% less likely to meet the criteria for suspected cheating (OR = 0.15, $p < .001$) and reported lower rates of self-reported online lookup (8.8% vs. 20.0%), mirroring a 3.6-fold difference in cheating rates (14.7% vs. 52.9%). The large effect size (Cramer's $V = .40$) underscores the practical significance of dialogic formats for integrity outcomes.

These results extend prior work showing high cheating prevalence in unproctored, recognition-based assessments—where file-share test banks and automated answer-seeking inflate scores by 13–18 percentage points (Lancaster & Cotarlan, 2021; Roediger & Marsh, 2005). Consistent with this literature, our multiple-choice participants scored approximately 20 percentage points higher than those in the dialogic condition, suggesting similar levels of artificial score inflation in the traditional format.

These findings align with theoretical and empirical literature emphasizing social accountability and process transparency as key integrity levers. Social-presence theory posits that richer media conveying nonverbal immediacy heighten perceptions of being "seen," thereby deterring dishonest behavior (Biocca, Harms, & Burgoon, 2003). The dialogic avatar's real-time, adaptive questioning established a conversational environment that mimicked human oral assessment, requiring learners to articulate and defend their reasoning in ways that multiple-choice formats cannot elicit. This interpersonal immediacy contrasts sharply with the anonymity of multiple-choice assessments, which weaken social norms and reduce the psychological cost of cheating (Holden, Norris, & Kuhlmeier, 2021; Ryan & Deci, 2020).

Beyond integrity benefits, dialogic assessments are likely to foster deeper learning through generative retrieval and tailored scaffolding. By requiring learners to articulate their conceptual understanding and respond to follow-up questions, dialogic formats situate assessment as a learning event "for" and "as" learning rather than merely "of" learning (Alexander, 2020; Vygotsky, 1978). These interactive dynamics capture rich process data—response revisions, latencies, and dialogue turns—that can serve as authenticity signals and diagnostic indicators of conceptual mastery.

Implications for practice and policy are noteworthy. Institutions should consider integrating AI-mediated dialogic assessments as part of a multifaceted integrity strategy that prioritizes design over detection (Evangelista, 2025). By embedding process transparency and requiring original articulation, educators can transform assessments into integrity-first experiences. Moreover, workforce-development programs that mirror professional verbal reasoning tasks—such as sales pitches, interviewing, and clinical presentations—stand to benefit from dialogic formats that both uphold fidelity to real-world practices and undermine external cheating opportunities. Ultimately, dialogic assessment offers a promising, scalable approach that preserves academic integrity in an era of ubiquitous generative AI.

References

Alexander, R. (2020). *A dialogic teaching companion*. Routledge. https://doi.org/10.4324/9781351040143.

BestColleges. (2023). *56% of college students have used AI on assignments or exams*. https://www.bestcolleges.com/research/most-college-students-have-used-ai-survey/.

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & Virtual Environments*, 12(5), 456–480.

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., van Haeringen, K., Saddiqui, S., & Rozenberg, P. (2019). Contract cheating and assessment design: Exploring the connection. *Studies in Higher Education*, 44(11), 1837–1856.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529), 86–89.

Common Sense Media, Hopelab, & Center for Digital Thriving. (2024). *Teen and young adult perspectives on generative AI*. https://hopelab.org/teen-young-adult-perspectives-generative-ai/.

Evangelista, E. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology*, 17(1), ep559.

Holden, O., Norris, M., & Kuhlmeier, V. (2021). Academic integrity in online assessment: A research review. *Frontiers in Education*, *6*, 639814.

Lancaster, T., & Cotarlan, C. (2021). Contract cheating by STEM students through a file sharing website: A COVID-19 pandemic perspective. *International Journal for Educational Integrity*, 17, Article 3, 1-16.

Roediger, H., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155–1159.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, Article 101860.

Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological science*, *28*(12), 1745-1762.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wiley Education. (2024). The latest insights into academic integrity: Instructor & student experiences, attitudes, and the impact of AI 2024 update. Lumina

Foundation. https://www.luminafoundation.org/wp-content/uploads/2024/08/The-Latest-Insights-into-Academic-Integrity.pdf.

Zaky, Y., & Gameil, A. (2024). Exploring the use of avatars in the sustainable Edu-Metaverse for an alternative assessment: Impact on tolerance. *Sustainability*, 16(15), 6604.

Author's Background

Kelly Puzio, Ph.D., is the Director of Research at Creatium. He is a learning scientist who received a Ph.D. in Learning, Teaching, and Diversity from Vanderbilt University, a Master's in Education from DePaul University, and a bachelor's degree from the University of Notre Dame. He was a certified teacher in New Zealand and the United States, and an Associate Professor at Washington State University in the Department of Teaching and Learning. His current research interests include personalized learning and artificial intelligence. Dr. Puzio has served as Principal Investigator on over $5 million in grants from the National Institutes of Health and the National Science Foundation. He was a recipient of the Young Scholar Award from the International Literacy Association.