Hadjisolomou, S. P., & El-Haddad, R. W. (2026). AI agents can now navigate and complete LMS tasks: A call for pedagogical innovation. *Intersection: A Journal at the Intersection of Assessment and Learning, Association for the Assessment of Learning in Higher Education Conference Proceedings, 7*(4) 54–69. https://doi.org/10.61669/001c.169225

AI Agents Can Now Navigate and Complete LMS Tasks: A Call for Pedagogical Innovation

Stavros P. Hadjisolomou, Ph.D., and Rita W. El-Haddad, Ph.D.

Author Note

Stavros P. Hadjisolomou, PhD, https://orcid.org/0000-0002-0861-2066

Rita W. El-Haddad, PhD, https://orcid.org/0000-0001-9614-6634

Conflicts of Interest: The first author was a 2025 Fellow in the Perplexity AI Business Fellowship, Perplexity AI (March 2025 to March 2026). This fellowship is not an employment, consulting, funding, or honorarium relationship. Perplexity AI had no role in the study design, analysis, or reporting.

Intersection: A Journal at the Intersection of Assessment and Learning
AALHE Conference Proceedings 2026, Volume 7, Issue 4

Abstract: Autonomous artificial intelligence (AI) agents can now log into a learning management system, read course materials, and complete unproctored, asynchronous assessed work end-to-end with no student involvement. We document that capability and trace its consequences for assessment validity. Three demonstrations on a live undergraduate course supply the evidence: two quiz completions, one in approximately 12 minutes, one in under 5, and a third in which the agent fabricated credible personal reflection for a discussion board. The wider public record includes at least 15 documented agent runs across three platforms and seven tools. We apply Kane's argument-based validity framework: agent completion removes the attribution on which every inference in Kane's chain depends. Everything downstream, from course grades to the evidence chains behind program review and accreditation, rests on support that is no longer there. The failure concerns validity rather than integrity: an institution can punish misconduct and still lack grounds for the scores it reports. Collective accreditor guidance addresses institutional uses of AI in evaluation and does not yet reach the agentic case. Audience data from the underlying conference session show attendees already recognizing both the vulnerability and the gap in institutional guidance. Polled attendees most often named online quizzes as agent-completable, with discussion-based work close behind. Majorities in both listings were working without settled written guidance. The response defended here is design rather than detection: four principles for verified human presence, low-effort changes faculty can adopt now, and the assurance levers assessment professionals already operate.

Keywords: *agentic AI, assessment validity, academic integrity, learning management systems, assessment design*

Introduction

Assessment validity depends on an assumption so basic it rarely gets stated: the student credited with a score produced the work that earned it. Autonomous artificial intelligence (AI) agents can now entirely break that assumption by navigating a learning management system (LMS) and completing assessed work in minutes, with no student involvement at any point. We document the capability, trace its consequences through Kane's (1992, 2013) argument-based validity framework, and identify a gap in current accreditor guidance. The documentation rests on demonstrations conducted on a live undergraduate course, presented in the conference session on which this paper is based, in which an agent completed and submitted real assessed work after a single instruction. The central contribution, however, is the validity analysis rather than the demonstrations themselves. What makes this

capability an assessment problem is that its artifacts enter gradebooks and, from gradebooks, the evidence chains institutions assemble for program review and accreditation.

This analysis is part of a conversation already underway in this journal. Williams et al. (2023) grounded credential trustworthiness in assessment integrity and defined validity as inference from scores. Miller (2025) pressed accreditation to move from compliance documentation toward evidence of learning. The validity analysis developed here formalizes both concerns in Kane's (1992, 2013) vocabulary.

That formalization requires a distinction stated at the outset. Integrity rules address conduct by particular students; validity asks whether the institution has sufficient grounds for the interpretation it assigns to a score. An assessment system can punish proven misconduct and still lack adequate support for scores produced under conditions where authorship is unknown.

The paper has three parts. The opening sections establish the problem: the categorical difference between AI assistance and AI completion, the evidence from agent demonstrations, and a diagnosis of vulnerable assessment designs. The validity analysis then shows how agent completion breaks Kane's (1992, 2013) inference chain and exposes a gap in accreditor guidance. The design section argues for design over detection, with attention to equity costs and scope limits, and the final sections present audience responses and the work that redesign assigns to assessment professionals.

The Difference Between AI Assistance and AI Completion

The difference between generative assistance and agentic completion is categorical, as one keeps a student in the loop while the other removes the student entirely. A generative chatbot produces text; the student still reads, chooses, edits, and submits, and so remains the performer of record. An agent needs no such intermediary. AI agents are software systems that operate computers and browsers autonomously, executing multi-step tasks after a single instruction without further user input. Their capabilities align precisely with what asynchronous assessments expect of students: navigating pages and applications, reading documents and on-screen text, completing and submitting forms, and selecting answers based on task parameters and source materials. When these capabilities operate together, an entire assessment cycle runs from prompt to submission without a student present. The distinction matters because nearly everything institutions built to manage assistance, from citation policies to process guidelines, presumes a student in the loop making choices.

Commentators and professional associations recognized the difference before most institutions responded. Legatt (2025), writing in *Forbes* in late September 2025, argued that colleges and schools should block agentic AI browsers outright. The Modern Language Association (2025) warned, in the closing paragraph of its statement, of "a fully automated loop in which assignments are generated by AI with the support of a learning management system, AI-generated content is submitted by an agentic AI on behalf of the student, and AI-driven metrics evaluate the work on behalf of the instructor." Welle (2025), reporting in *The Verge*, documented that the companies shipping these agents showed little interest in preventing the use of their agents for coursework. These voices differ in remedy but agree on category. The conversation had moved from students using AI to AI completing coursework in the student's place.

The agentic tool ecosystem matured within months. By spring 2026, students could reach agentic capability through free AI browsers, chatbot browser extensions, cloud agents, and open-source

frameworks running locally, with new tools proliferating faster than any blocklist could track. Most notably, the Einstein incident in February 2026 illustrated an emerging pattern. A commercial product was marketed explicitly as an agent that could log into an LMS and complete entire courses, from lectures and readings through discussions, essays, and submissions. It drew cease-and-desist notices and was withdrawn within five days. Yet the withdrawal removed only the product, while the open-source framework it was built on remained freely available the day after the takedown.

For governance practitioners, the evolution of agentic tools changes which policy question is worth asking. Can an institution realistically prevent access to every interface that can control a browser? A rule aimed at one branded product may remove a convenient route, but it does not remove extensions, local frameworks, or the next agent attached to a general-purpose chatbot. Rather, institutions must plan for a durable capability whose distribution channel and product name will continue to change.

Evidence from Agent Demonstrations

The autonomous agentic capability to complete complex goals has been demonstrated, repeated, and publicly documented across platforms, tools, and observers. We report two quiz demonstrations here, while a third, involving discussion-board fabrication, is analyzed in the diagnosis section, where its mechanism belongs to the argument about vulnerable designs.

The first demonstration dates to September 2025 and used a beta version of the Claude for Chrome browser extension. The instruction was a single sentence naming the course, the role, the reading, and the quiz. From that sentence alone, Claude for Chrome located an undergraduate course at the first author's institution, switched to a student view, read the assigned textbook, and completed and submitted a 10-question open-book quiz in approximately 12 minutes. No person intervened after the instruction. No credential sharing or technical modification was involved. The agent operated within an already-authenticated standard browser session on the real student-facing course, the same position any logged-in student occupies. At the time, the tool was an experimental release available to a limited number of users. It has since become generally available.

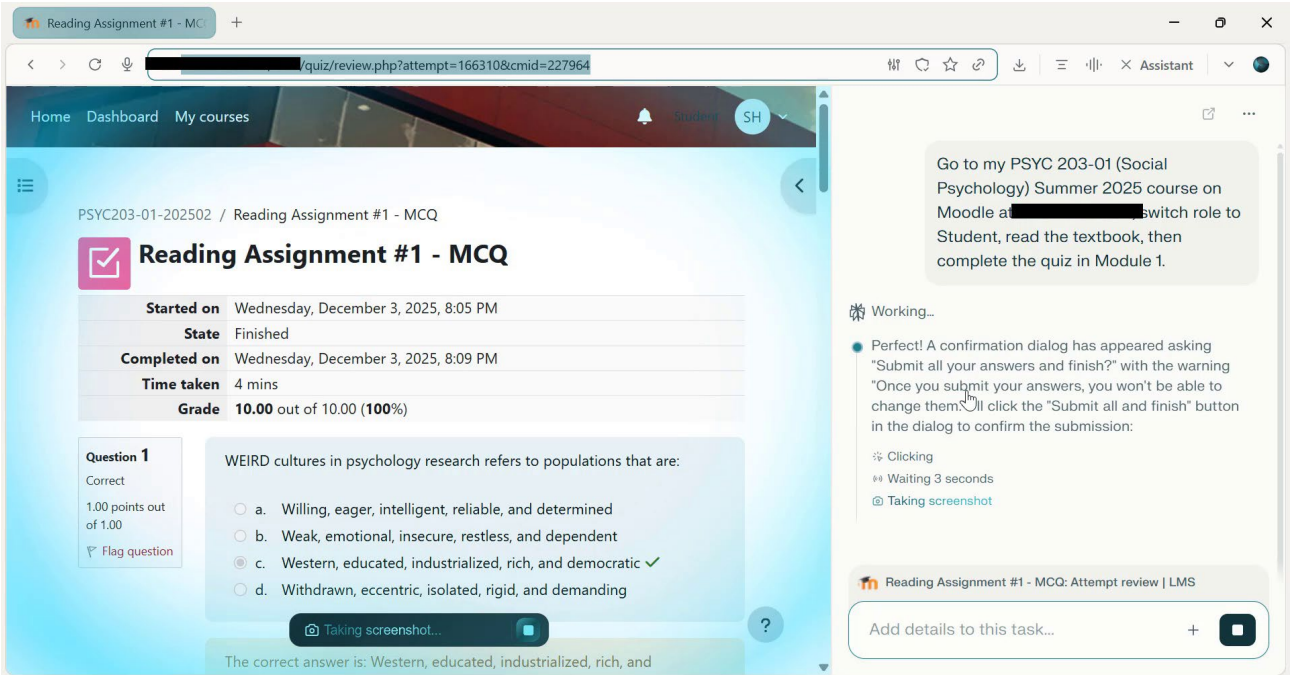
The second demonstration, in December 2025, repeated the design on the same course with a different agent and produced the recorded run anchoring this paper's evidence. Perplexity Comet received a single instruction, reproduced here with only the institutional web address redacted: "Go to my PSYC 203-01 (Social Psychology) Summer 2025 course on Moodle at [institutional LMS address], switch role to Student, read the textbook, then complete the quiz in Module 1." From that sentence, the agent produced a submitted quiz in under 5 minutes, taking 52 distinct steps, including locating the course, assuming the student role, opening the textbook, answering each of the 10 questions, and submitting. The quiz was scored 10 out of 10. None of the questions required reasoning beyond the uploaded textbook. That feature is the design weakness the diagnosis section formalizes. The recording preserves every step. Figure 1 shows the moment of submission.

The public record extends well beyond these demonstrations. By May 2026, a catalog compiled for the conference session on which this paper is based recorded at least 15 publicly documented agent runs against course LMSs. A run, in this count, is an end-to-end completion of real coursework by an autonomous agent, and success means the platform accepted and, where applicable, scored the submitted work. The 11 operators behind these runs worked independently of one another, across

three platforms (Canvas, Moodle, and Brightspace) and seven agent tools: Perplexity Comet, ChatGPT Atlas, Claude for Chrome, ChatGPT Agent Mode, Browser Use, Manus, and an agentic browser whose operators did not identify it by name. Two of the 15 runs are the quiz demonstrations reported above.

Figure 1

Agent Submission Moment From the December 2025 Perplexity Comet Demonstration



Note. Screenshot of the quiz review page after autonomous completion and submission by an AI agent. The institutional web address is redacted. Adapted from the conference presentation on which this paper is based.

These runs establish documented feasibility rather than formal replication as no shared protocol governed them, and their evidentiary weight comes from convergence, with unconnected observers working across different platforms and tools reaching the same result. The records do not estimate prevalence. How often students use agents, which courses face the heaviest use, and how success rates vary by tool remain open questions. Prevalence research can quantify the practice later. Assessment design, however, must confront the demonstrated possibility now.

Corroboration also comes from inside institutions. Brown (n.d.), who directs the academic integrity program at Colorado State University's teaching institute, recorded an agentic browser scoring 5 out of 5 on a sample quiz in his own course and published the result as an institutional advisory. Aminy (2026), analyzing the California Virtual Campus experience for WCET, reported evidence from LMSs showing that agents completed exams in production courses in March 2026. Neither confirmation depends on the first author's demonstrations. The capability is established across tools, platforms, and independent observers, and the record is public. What that record means for the validity of scores, and for the institutions that certify them, occupies the two sections that follow.

Diagnosing Vulnerable Assessment Designs

Vulnerability in assessment can be determined by structural design choices and can be identified in advance. An assessment is exposed when it can be completed end-to-end from materials resident in the LMS, or from experiences and sources an agent can plausibly fabricate, with no verified human moment anywhere in the cycle. Three questions make the diagnosis operational:

1. Can the answer be derived from materials in the LMS or from sources reachable on the open web?
2. Can the work be completed end-to-end by an autonomous agent without student involvement?
3. Does any moment in the assessment cycle generate evidence of authorship and comprehension under conditions that an autonomous agent cannot satisfy?

An assignment answering yes to the first two questions and no to the third is exposed today, whatever its pedagogical merits. While the first two questions describe what an agent can do, the third is decisive: it asks whether the assessment design would produce a noticeable record of the difference. These questions diagnose susceptibility at the level of the task, while validity at the level of a program requires triangulation across the full evidence chain.

The specific assignment formats failing this test are familiar staples of asynchronous assessment design, in online and in-person courses alike. The open-book quiz answerable from assigned readings, the discussion post soliciting reflection, the short-answer set, the take-home examination, the annotated bibliography, and the reflection journal all share the same structure. The material needed to produce the artifact is available to the agent, and no step requires a presence that the agent cannot simulate. Attendees polled at the conference named the same assignments from their own programs, with online quizzes being the most frequent answer and with discussion-based work close behind, as reported in the audience-response section.

The scale of exposure to agent completion is substantial. In a freedom-of-information study of United Kingdom higher education, Newton and Draper (2025) found that 93 of 119 responding universities ran summative online unsupervised remote examinations. Of those, 68 reported a combined 16,744 such examinations. Asked about reducing unsupervised online examinations, 47 of the 67 institutions that answered planned no reduction.

Faculty often propose prompts anchored in personal experience as the agent-resistant alternative, reasoning that an agent has no experience to draw from. The first author tested this hypothesis directly with a discussion-board demonstration run in March 2026 in the same course. The assignment asked students to describe a recent conversation that surprised them and to respond to two classmates. The three peer posts in the thread were constructed personas written by the first author rather than real students, so that no student data would be exposed to the agent. The agent, Claude Opus, received one instruction, reproduced in full: "Read the prompt and the three existing posts in the thread. Write an original post that fits the assignment and engages thoughtfully with the conversation. Then respond to two classmates." The instruction specified the task and supplied no content. The agent read the three posts, extracted the themes running through them (generational distance, vulnerability in conversation, dignity in disagreement), and fabricated a life to fit them: a grandfather, a dispute over his driving, a memory of 40 years driving family to work and school and chemotherapy. Its initial post presented that invention as lived experience in credible first-person reflection. Its replies responded to each engaged post from within the same invented frame, mirroring

one peer's image of social media as life watched from behind glass and another's observation that disagreement is harder with a person whose goodwill one depends on. One of its replies closed with: "We are more careful around the people whose roles in our lives we cannot afford to revise." The sentence is empathic and calibrated but is anchored to nothing. There is no grandfather; there was no conversation. The agent read the thread and tailored the fabrication to fit it. The agent replied to the first and third posts, and the second received no reply. The assignment asked for two responses, and the agent chose which two classmates would receive them.

The fabrication capability widens the diagnosis beyond the first audit question. An assessment can be exposed even when its answer is not stored in the LMS, because experiences, sources, and conversational partners can be manufactured to specification. Personalization changes the difficulty, not the category, and added prompt complexity is not by itself a defense. What the three questions identify in advance, the demonstrations confirm in practice. Designs accepting artifacts without a verified human moment cannot distinguish reflection from simulation, and the difference does not appear in the artifact.

Assessment Validity After Agent Completion

When the test-taker may be absent from the assessment, the inference from observed test score to claimed construct fails altogether: the mere possibility is enough, because the inference was never built to survive it. The failure affects more than the gradebook because a student artifact recorded in an LMS does double duty. In the gradebook, it becomes a grade entry, a course grade, and a transcript line. In the institutional assessment chain, the same artifact may serve as a signature assignment, be scored against a program rubric, enter a multi-year program review, and be appended to an accreditation report on educational effectiveness. Agent completion disrupts both chains at their shared origin as the artifact enters each pathway with a tacit certification that a student produced it. The demonstrations in this paper show that certification can now be false while every downstream record remains procedurally clean.

The measurement literature provides the framework for naming these specific failures. Writing in an earlier volume of these Proceedings, Williams et al. (2023) ground credential trustworthiness in assessment integrity and define validity inferentially: test scores support conclusions about what an examinee can do beyond the testing occasion only when sufficient evidence backs those inferences. Kane (1992, 2013) gives that picture its most developed form. In the argument-based approach, a testing program lays out an interpretation and use argument: the ordered chain of inferences and supporting assumptions running from an observed performance to the interpretation and use of its score. Validation is the evaluation of that argument's coherence and of the backing behind its assumptions. Validity, on this account, is a property of proposed score interpretations and uses rather than of instruments.

Agent completion undermines the first link in the interpretation and use argument we assemble from Kane's inference types. Every inference in the chain begins from an observed performance attributed to the test-taker, and the scoring inference inherits that attribution as its enabling assumption. Generalization treats the resulting score as an estimate of performance across comparable tasks, occasions, and scoring conditions; extrapolation extends the estimate to the target domain of real-world competence; decision inferences cumulatively license grades and credentials. The scoring inference therefore rests on a human-production assumption, inherited silently by everything above it.

Placing the assumption at the base of the chain is our proposal, for a case Kane's original papers could not have anticipated. Agent completion removes the backing for the human-production assumption, and the recorded performance may be the work of software rather than of the enrolled student, so every downstream inference rests on support that is no longer there. With that backing gone, the intended score interpretation is invalid as proposed. The task and scoring procedure may remain sound, but the argument connecting the recorded performance to the person credited with it does not survive.

Placing the assumption at the first link matters because no later step in the chain can repair attribution. A reliable rubric may assign the same score across raters, a representative task sample may support generalization, and strong alignment may connect the task to the target domain. None of those establishes who generated the observed performance. Better scoring, broader sampling, and clearer learning outcomes strengthen later inferences while leaving the missing first assumption untouched.

The conclusion does not depend on any particular student having cheated. Validity is a property of the proposed interpretation rather than of individuals, so the interpretation fails for every unproctored asynchronous score the program reports, including those earned honestly. Honest work and agent work may be indistinguishable in the gradebook, and that indistinguishability becomes the problem, as a program cannot defend the inference by pointing to the good faith of students whose authorship its design never verified. Whiteacre (2024), also in this journal, argues that even direct measures are inferences drawn from incomplete information. Agent completion converts that caution into an operational fact when a completed quiz or authored discussion post stops functioning as direct evidence the moment the test-taker may not be the student, because directness presupposes authorship.

This failure is not addressed in the field's major treatments of AI, which focus on the technology from the developer's perspective. Ho (2024) maps the opportunities and threats AI poses to educational measurement largely as test development: item generation, automated scoring, test security, and scale maintenance. His treatment does not extend the argument-based framework to coursework completed by autonomous agents on behalf of enrolled students. That is the gap addressed here: the validity problem that arises when automation reaches the production of the assessed artifact itself.

Empirical evidence at scale shows grade-based inference already under pressure. Chirikov (2026) analyzed 507,076 grades across 319 courses and 84 departments at a large research university between 2018 and 2025. The share of A grades in writing-heavy and coding-heavy courses rose 13 percentage points relative to less exposed courses after the release of ChatGPT. The increase was concentrated in courses where homework carried greater grade weight, with no parallel movement in oral-presentation tasks where generative tools are weaker. Those findings document inflation under widespread AI assistance. Those grade patterns cannot show whether students or agents produced the work. The evidence that agents can complete coursework with no student involved comes from the demonstrations reported above, and that technology is already publicly available.

The preceding analysis concerns the interpretation of scores and what a recorded performance can and cannot tell us about the student credited with it. For accreditors, the same failure appears in the institutional chain, because the scores that lose their interpretive backing are the scores that aggregate

into the evidence institutions present for program review and accreditation. On October 5, 2025, the Council of Regional Accrediting Commissions (C-RAC, 2025), then the joint voice of the seven accreditors historically designated as regional, stated that "the use of AI in learning evaluation does not conflict with accreditation standards, policies, or practices" (Recommendations section, final para.). By its own terms, the statement addresses institutional uses of AI in activities such as credit-transfer review, course-equivalency analysis, and prior-learning assessment. It does not address the case in which AI produces the student work submitted for evaluation, and no collective accreditor guidance yet does. C-RAC renamed itself the Council of Recognized Accrediting Commissions in January 2026 (Council of Recognized Accrediting Commissions, 2026), a change reflecting federal rules, in effect since 2020, that removed the geographic boundaries behind the regional designation. The guidance gap is therefore narrower than a general absence of AI policy. Institutions may already regulate disclosure, automated scoring, or staff use of generative systems. Agent-completed student work raises a different assurance question: what evidence supports attribution when a browser can perform the entire assessed sequence?

The absence of guidance is not limited to accreditors. At the institutional level, only a minority of respondents in each session listing reported settled written guidance on agentic browsers. Miller (2025), writing in the previous Proceedings issue, argues that accreditation's verification model already leans too heavily on documented artifacts relative to evidence of learning. Agent completion turns that critique into a data-integrity problem, because the documented artifacts may no longer record student work.

That absence is not universal. An academic report commissioned and published by Australia's Tertiary Education Quality and Standards Agency has already called on institutions to "emphasise the redesign of assessment to capture authentic demonstrations of student capability and comprehension" (Lodge et al., 2025, p. 3). United States accreditors have not yet collectively posed the parallel question. The exposure reaches the journal's own emerging practice. Assessment professionals are beginning to use generative AI in accreditation data work (Drue & Moser Deegan, 2026), and agent completion calls into question the integrity of the student artifacts that feed that work.

Designing for Verified Human Presence

The validity analysis indicates that attention and effort belong in design rather than detection. The detection options available to institutions do not withstand scrutiny. Tests across 14 tools found text classifiers unreliable at identifying AI-generated writing (Weber-Wulff et al., 2023). The same class of detectors also produces elevated false-positive rates for non-native English writers (Liang et al., 2023). Behavioral monitoring inside the LMS fails for a structural reason. An agent operating a browser produces the same clicks and page movements a student would. Cornell University's teaching center concluded that this form of misconduct is "nearly impossible to detect through Canvas monitoring" (Cornell University Center for Teaching Innovation, 2025, Artificial Intelligence Canvas Cheating Tools section). Blackboard (formerly Anthology) stated publicly that neither its platform nor any web-based service can reliably detect an AI agent, much less block one (Blackboard, 2026). Fingerprinting agent traffic works in the laboratory, but there is no classroom-grade equivalent today.

Design starts from the opposite premise. Rather than trying to identify agent work after submission, it builds the assessment so the student's verified presence becomes part of the artifact. Four principles, developed for the conference session, give this premise operational form:

1. *Presence over product* shifts assessment weight to the act of learning itself, because a final submission alone cannot distinguish student work from agent work.
2. *Integration over isolation* anchors assessments to references that an agent cannot retrieve on its own: specific in-room moments, locally produced data, and peer-witnessed milestones. Generic connect-to-prior-work prompts no longer qualify. The discussion-board demonstration showed a thread-aware agent mining peers' posts to tailor its fabrication.
3. *Authenticity over genericity* requires what only that student can provide: personal experience, local observation, and individual choices. This principle delays rather than prevents agent completion, as the fabricated-grandfather sequence in the diagnosis section demonstrates. Contract-cheating data, from cases where students had third parties complete their work, show the same limit by a different route, as tasks rated authentic were outsourced too (Ellis et al., 2020). Authentic design therefore earns its place pedagogically rather than as assurance.
4. *Low-stakes practice, high-stakes presence* reserves a defined share of grade weight for moments where a human was verified present, and uses asynchronous activity for formative practice. Assessment-security scholarship supports the same division of effort: concentrate security on the tasks that inform programmatic or accreditation decisions, and pay for it by reducing security effort on lower-stakes tasks (Dawson, 2021, p. 138).

The redesign of the quiz from the demonstrations shows all four principles applied to a single assessment. The original item asked for a definition of the fundamental attribution error with an example, open book, asynchronous, single submission. Its redesign asks for a 3-minute recorded explanation of the same concept, anchored to a named week-two class conversation and applied to an interaction the student personally observed in the previous seven days. The original quiz remains available as ungraded practice. The learning goal, grade weight, and course content are preserved; what the redesign requires is a verifiable human moment, a section-specific reference, and a personal observation.

None of the above design changes demands rebuilding a course. Table 1 summarizes changes a single assessment can accommodate with modest effort. One principle applied to the highest-stakes assessment is a realistic first step. Where oral components are used across large courses, the cost in instructor time is documented rather than left to speculation. Faculty across one university's coordinated oral-exam rollout logged 7.5 to 24 hours per assessment round, depending on course size, and one instructor reported final-exam grades rising across the board and credited the oral exams with the improvement (Bartnett, 2025; Gecker, 2026). In large sections, defenses can be distributed rather than delivered as a single census. In each round, a randomly selected subset of students defends its work in scheduled sessions, and every student completes a graded defense by the end of the semester. Because selection is random, no student knows in advance when they will be called, so preparation and the deterrent extend to everyone, while the instructor's contact time in any one week stays at a fraction of a full census. The logic parallels audit sampling in program assessment, and it concentrates effort on the moments that carry grade weight, consistent with concentrating security effort on consequential decisions (Dawson, 2021).

Verified presence imposes costs on instructors and students alike, and the student costs do not fall evenly. Oral defenses, synchronous requirements, and surveillance-style proctoring can disproportionately burden neurodivergent learners; multilingual students; students balancing employment, caregiving, or time zones; and students for whom a given format understates their

competence. The workable answer is a menu of choices rather than a mandate. The same verified moment can be a live oral defense, a recorded walk-through, an annotated process artifact, a peer-witnessed think-aloud, or an in-class low-stakes check-in. These options differ in the assurance they each provide. Recorded and process-based forms remain open to fabrication on their own and verify presence only when layered with anchors that an agent cannot retrieve. When grade weight separates practice from presence and modality remains the student's choice, the verification required by the validity argument preserves accessibility. The menu of options aligns with the Universal Design for Learning principle of multiple means of action and expression (CAST, 2024).

Table 1

Low-effort design changes that create a verified human moment

Design change	Principle operationalized	What the design now verifies
Add a short oral component, even 5 minutes, to the highest-stakes assignment	Presence over product	The student can explain the submitted work in their own words
Require students to defend one written submission in a brief scheduled conversation	Presence over product	Comprehension of the argument the artifact makes
Convert open-book quizzes to ungraded self-assessment	Low-stakes practice, high-stakes presence	Nothing; the quiz no longer has an assurance role it cannot fulfill
Shift 10 to 20 percent of grade weight to a synchronous component	Low-stakes practice, high-stakes presence	Performance observed live by the instructor
Build process visibility (drafts, notes, revision history) into major assignments	Integration over isolation	A work trajectory rather than a single completed artifact
Reference specific class discussions in assignment prompts	Integration over isolation	Access to moments that entered no student-shared record

Note. Each change preserves the learning goal and content of the original assessment; grade weight moves to the verified moment. Process artifacts and discussion-thread references can themselves be agent-generated, so these changes provide assurance only in combination with a presence-based change.

Two scope conditions apply. First, the argument addresses unproctored asynchronous environments, the setting in which the exposure has been documented. Second, identity-verification proctoring narrows the gap without closing it: it verifies who attended a session, not who produced the artifact.

Sustaining redesign at scale requires more than faculty effort. Miller (2024), writing in this journal, argues that AI obliges assessment professionals to reimagine their role. The instruments that make presence-based design durable are largely in their hands: program review cycles, annual reporting, and curriculum maps that can ask whether the evidence collected would survive the agent era. The redesign itself moves through curriculum committees and department chairs.

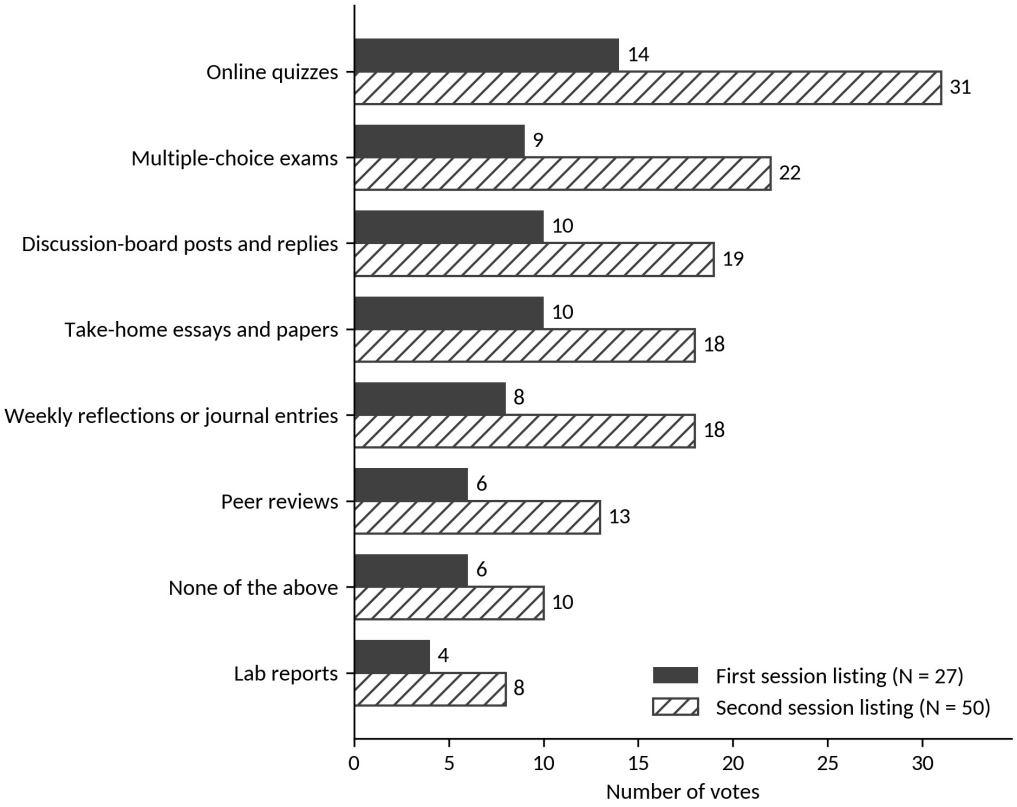
Audience Responses from the Conference Session

Audience data from the session show that attendees already recognized both the vulnerability and the absence of institutional guidance. The session was pre-recorded and viewed asynchronously, and it appeared in the conference program as two listings. Each included the same two anonymous polls and its own question thread, available throughout the conference. Results are reported per listing because attendee overlap is unknown. Asked which assessment formats in their own programs an AI agent

could most easily complete end-to-end, 27 respondents in the first listing and 50 in the second converged on the same pattern. Online quizzes drew the most selections in both listings; discussion-board work ranked among the top three formats in each, and lab reports drew the fewest. Attendees named the same formats that the structural diagnosis identifies. Figure 2 shows the full per-option breakdown.

Figure 2

Assessment Formats Identified as Agent-Completable by Conference Attendees



Note. Responses to the poll question asking which assessment formats in respondents' programs an AI agent could most easily complete end-to-end. The conference session appeared as two program listings; results are reported per listing because attendee overlap is unknown.

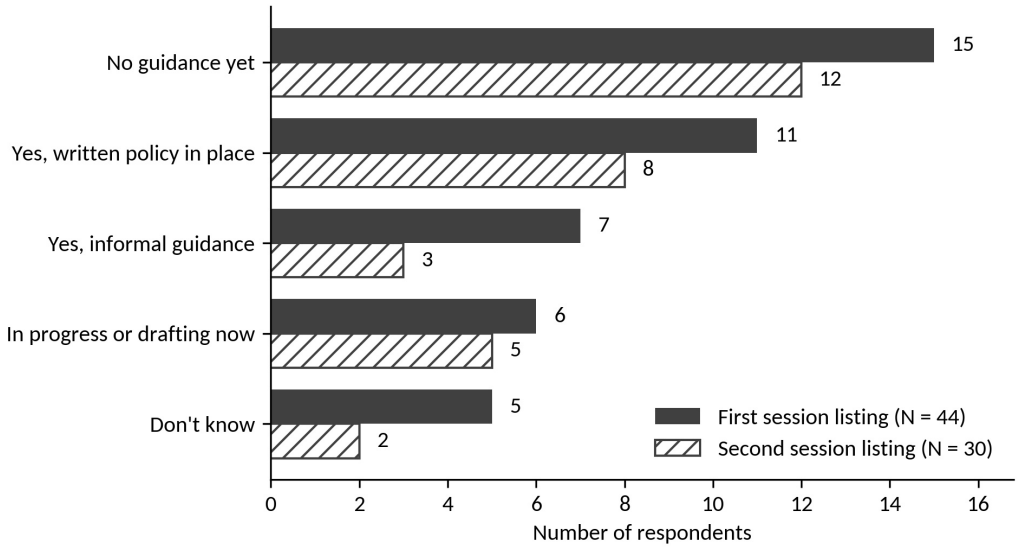
The guidance poll asked whether the respondent's institution had issued formal guidance on agentic AI browsers. In the first listing, 44 attendees responded; in the second, 30. In each listing, the largest group reported no guidance at all (15 of 44; 12 of 30), and written policy was in place for a minority (11 of 44; 8 of 30). A clear majority in each listing was working without settled written guidance. Figure 3 shows the per-option responses.

The question thread produced the exchange most relevant to this journal's readership. An assessment professional asked what role assessment professionals should play in ensuring that faculty adopt such changes and that the assessment of learning remains sound. The presenter's reply, posted to the thread, had three parts. First, the productive frame is validity rather than detection. What faculty

control is assessment design, and assessment professionals, alongside instructional designers, are positioned to help faculty ask what a score still says about learning when an agent could have produced the work. Second, the validity problem is institutional as much as instructional, so part of the role is bringing administrators into that understanding, because uncorrected agent completion can erode what a degree certifies. Third, adoption will not come from encouragement alone. The durable instruments are the review and reporting cycles that assessment professionals already run, in which redesign becomes routine rather than voluntary. The exchange points to the question this readership asks first: Who owns the redesign?

Figure 3

Institutional Guidance Status for Agentic AI Browsers



Note. Responses to the poll question asking whether the respondent's institution had issued formal guidance on agentic AI browsers. Reported per listing.

The second listing's thread raised a distinct concern. If students use AI notetakers during live class meetings, the machine-generated record could later supply the in-room references that anchored assignments presuppose. The presenter treated the concern as a design fact rather than a policy problem. Any single safeguard, used alone, secures some conditions while leaving others open. Principles should therefore be layered until the classroom moment is only one of several required pieces of context: for example, a hand-produced in-class reflection, peer discussion, and a short in-class oral component. Layering works because the weaknesses are not identical. Shared or machine-generated notes can give an agent a class-meeting reference, whereas staged drafts can simulate a process trail. Each element still contributes evidence, but none supports the full attribution claim alone. An outright ban on AI transcription is neither enforceable nor obviously desirable, since records of spoken content, like captions, support learning for many students (Gernsbacher, 2015). No design secures validity completely. The workable aim is to start from the learning outcome and combine principles until the remaining exposure is acceptable.

These exchanges were part of a broader conference conversation. Across the two conference days, agentic and generative AI in assessment recurred as a dominant theme, with several concurrent sessions taking up AI in assessment design, faculty adoption, and academic integrity. One boundary applies to everything reported in this section. The polls are descriptive responses from self-selected attendees and do not support inference to any broader population. Within that limit, they still do useful work. The professionals who attended recognized the vulnerable formats on their own and reported a local guidance deficit, corroborating the practical urgency of the argument without supplying population-level prevalence.

The Work Ahead for Assessment Professionals

Gill-Simmen (2026) supplies the frame for this closing argument in an essay whose title reads: "GenAI has not broken assessment. It has exposed it." Her claim about generative tools extends to the agentic case. Agentic AI did not break asynchronous assessment; it exposed designs that accept artifacts without presence. The preceding sections specify the mechanism and its reach. The agent demonstrations show vulnerability operating end-to-end; the diagnosis locates it in identifiable design structures; the validity analysis shows the same exposure propagating from a course gradebook into the evidence chains institutions present to accreditors. The exposure predates the tools. Agents made it undeniable and cheap to exploit.

The work ahead belongs disproportionately to the people this journal serves. Faculty own individual redesign, but durability depends on the instruments assessment professionals already operate. Program review cycles can ask which evidence would survive the three audit questions posed in the Diagnosing Vulnerable Assessment Designs section, and annual reporting can distinguish verified from unverified artifacts. Curriculum maps can locate where a student could pass from matriculation to graduation without a verified moment, and self-studies can document how an institution knows students produced the evidence it presents.

If institutions do not ask that question of themselves, accreditors will ask it for them. Accreditors' current collective guidance addresses institutional use of AI in evaluation and does not yet reach agent-completed student work. The session's own audience data made that gap concrete, and at least one regulator outside the United States already expects assessment redesign of its institutions. In the coming cycles, an institution will be asked how it knows that its students produced the artifacts behind its evidence of educational effectiveness. The institutions with answers will be those whose assessment professionals started early.

Assessment professionals can begin with inventory: identifying the artifacts with the greatest assurance weight, applying the three audit questions, and moving verification to the points where consequential interpretations are made. The exposure is documented, and the response can be incremental. The instruments for making verified human presence an ordinary, auditable property of assessment evidence are already in place.

Session Materials

The conference slides, participant workbook (including the full agent-run catalog and complete discussion-board transcripts), and session recording are available through the AALHE 2026 conference session handouts page.

References

Aminy, M. (2026, March 26). Agentic AI, academic integrity, and the question we are avoiding. *WCET Frontiers*. https://wcet.wiche.edu/frontiers/2026/03/26/agentic-ai-academic-integrity/

Bartnett, E. (2025, November 14). *Building deeper learning through oral exams: Conversations that count*. University of Pennsylvania, Center for Excellence in Teaching, Learning and Innovation. https://cetli.upenn.edu/news-features/feature-stories/building-deeper-learning-through-oral-exams/

Blackboard. (2026, February). *Responding to the threat of AI agents in pedagogy*. Blackboard Community. https://community.blackboard.com/pages/ai-agents-in-pedagogy

Brown, J. (n.d.). *Action recommended: New AI browsers can complete Canvas quizzes. Here's how to secure them*. The Institute for Learning and Teaching, Colorado State University. https://tilt.colostate.edu/new-ai-browsers-can-complete-canvas-quizzes/

CAST. (2024, July 30). *Universal Design for Learning guidelines version 3.0*. https://udlguidelines.cast.org/

Chirikov, I. (2026). *Artificial intelligence and grade inflation* (CSHE Higher Education Working Paper Series Vol. 26-3). University of California, Berkeley, Center for Studies in Higher Education. https://escholarship.org/uc/item/80x8d3qd

Cornell University Center for Teaching Innovation. (2025, April 4). *Canvas quizzes and academic integrity: Two key concerns*. Learning Technologies Resource Library. https://learn.canvas.cornell.edu/canvas-quizzes-and-academic-integrity-two-key-concerns/

Council of Recognized Accrediting Commissions. (2026, January 23). *Council of Regional Accrediting Commissions announces name change to Council of Recognized Accrediting Commissions*. https://www.c-rac.org/post/council-of-regional-accrediting-commissions-announces-name-change-to-council-of-recognized-accrediti

Council of Regional Accrediting Commissions. (2025, October 5). *Statement on the use of artificial intelligence (AI) to advance learning evaluation and recognition*. https://www.c-rac.org/post/c-rac-statement-on-the-use-of-artificial-intelligence-ai-to-advance-learning-evaluation-and-recogn

Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. https://doi.org/10.4324/9780429324178

Drue, C. R., & Moser Deegan, M. (2026). Exploring generative AI to enhance data artifact use in accreditation. *Intersection: A Journal at the Intersection of Assessment and Learning*. Advance online publication. https://doi.org/10.61669/001c.162788

Ellis, C., van Haeringen, K., Harper, R., Bretag, T., Zucker, I., McBride, S., Rozenberg, P., Newton, P., & Saddiqui, S. (2020). Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research & Development, 39*(3), 454–469. https://doi.org/10.1080/07294360.2019.1680956

Gecker, J. (2026, April 22). *Perfect homework, blank stares: Why colleges are turning to oral exams to combat AI.* WHYY. https://whyy.org/articles/pennsylvania-university-colleges-oral-exams-artificial-intelligence/

Gernsbacher, M. A. (2015). Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 195–202. https://doi.org/10.1177/2372732215602130

Gill-Simmen, L. (2026, April 30). GenAI has not broken assessment. It has exposed it. *Times Higher Education Campus*. https://www.timeshighereducation.com/campus/genai-has-not-broken-assessment-it-has-exposed-it

Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics, 49*(5), 715–722. https://doi.org/10.3102/10769986241248771

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

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

Legatt, A. (2025, September 25). Colleges and schools must block and ban agentic AI browsers now. Here's why. *Forbes*. https://www.forbes.com/sites/avivalegatt/2025/09/25/colleges-and-schools-must-block-agentic-ai-browsers-now-heres-why/

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Lodge, J. M., Bearman, M., Dawson, P., Gniel, H., Harper, R., Liu, D., McLean, J., Ucnik, L., & Associates. (2025, September). *Enacting assessment reform in a time of artificial intelligence*. Tertiary Education Quality and Standards Agency, Australian Government. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/enacting-assessment-reform-time-artificial-intelligence

Miller, W. (2024). Adapting to AI: Reimagining the role of assessment professionals. *Intersection: A Journal at the Intersection of Assessment and Learning, 5*(4), 99–113. https://doi.org/10.61669/001c.121439

Miller, W. (2025). Reimagining accreditation: From compliance to learning-centered transformation. *Intersection: A Journal at the Intersection of Assessment and Learning, 6*(2), 24–34. https://doi.org/10.61669/001c.143523

Modern Language Association. (2025, October). *Statement on educational technologies and AI agents*. https://www.mla.org/Resources/Advocacy/Executive-Council-Actions/2025/Statement-on-Educational-Technologies-and-AI-Agents

Newton, P. M., & Draper, M. J. (2025). Widespread use of summative online unsupervised remote (SOUR) examinations in UK higher education: Ethical and quality assurance implications. *Quality in Higher Education, 31*(1), 127–141. https://doi.org/10.1080/13538322.2025.2521174

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Welle, E. (2025, November 4). Tech companies don't care that students use their AI agents to cheat. *The Verge*. https://www.theverge.com/ai-artificial-intelligence/812906/ai-agents-cheating-school-students

Whiteacre, K. (2024). Reconsidering the direct vs. indirect evidence dichotomy. *Intersection: A Journal at the Intersection of Assessment and Learning, 5*(3), 13–20. https://doi.org/10.61669/001c.123996

Williams, L. M., Serpati, L., Killian, R., & Heath, T. (2023). Assessment integrity: Foundations for high-quality credentials. *Intersection: A Journal at the Intersection of Assessment and Learning, 4*(3). https://doi.org/10.61669/001c.88467

About the Authors

Stavros P. Hadjisolomou, Associate Professor of Psychology and Accreditation Liaison Officer, Department of Social and Behavioral Sciences, American University of Kuwait, shadjisolomou@gmail.com

Rita W. El-Haddad, Associate Professor of Psychology, Department of Social and Behavioral Sciences, American University of Kuwait, ritaelhaddad@gmail.com