Journal of Learning and Development Studies
ISSN: 2752-9541
DOI: 10.32996/jlds
Journal Homepage: www.al-kindipublisher.com/index.php/jlds

JLDS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Beyond AI-Proofing: A Conceptual Framework for Assessment Design in the Age of Generative AI

Bochra BOUDJEMAA
Lecturer, School of Humanities, Arts & Applied Sciences, Amity University Dubai, United Arab Emirates.
Corresponding Author: Bochra BOUDJEMAA **E-mail**: boudjemaabochra@gmail.com

| ABSTRACT

Though generative artificial intelligence now shapes many academic policies, institutions often react through bans, surveillance, and rigid rules. These measures fail at core levels, as they demand obedience to standards that lack practical enforceability, producing mere appearances of control instead of real educational integrity. This study suggests a different path under the name *"Beyond AI-Proofing"* that offers a restructured approach to assessment redesign built on four teaching and learning pillars. The first is validity: assessment should be about being confident that students can do what their degree says they can, not about searching for discrepancies that indicate potential academic dishonesty. The second is evaluating judgment: students need to learn to judge the quality of their work, other people's work, and AI's work, and this is something no machine can do on their behalf. The third is authentic process-focussed assessment: tasks should reflect real-world work and make the learning process visible, while being honest that authenticity is not a magic shield against AI. The fourth is a two-lane program structure: some assessments run securely in person, while the rest are open and AI is used openly with the whole program rather than any single task carrying the weight of proof. What ties these four together is that they build trustworthiness into the design of assessment itself, rather than trying to bolt it on through rules. Such a system holds up whether or not students use AI, because its integrity comes from how it is built.

| KEYWORDS

Assessment design, generative artificial intelligence, assessment validity, evaluative judgment, authentic assessment, academic integrity.

| ARTICLE INFORMATION

ACCEPTED: 09 June 2026 **PUBLISHED:** 30 June 2026 **DOI:** 10.32996/jlds.2026.6.8.3

1. Introduction

Generative artificial intelligence (Gen-AI) tools that produce convincing text, code, and analysis from a simple prompt have landed in higher education fast. As a reaction, universities responded just as quickly: traffic-light policies, AI assessment scales, detection software, declaration forms, and a return to in-person exams, all appeared within the same aim that is to stop students from using AI to complete work without actually learning anything. The worry behind this is fair, if students can hand their assignments to AI and still graduate, then the degree no longer certifies what it claims to, and assessment loses its whole point (Dawson et al., 2024). However, almost all of these responses share the same underlying logic, they treat AI as a discipline problem, and as a solution they simply set clear rules, then make students obey them.

A growing body of research shows that this approach simply does not hold up. AI-generated texts cannot be reliably detected (Corbin et al., 2025a), and detection software falls hardest on non-native English speakers, who are flagged wrongly far more often (Perkins et al., 2024). The line between getting acceptable help from AI and improperly handing the work over is genuinely blurry, for students and teachers alike (Corbin et al., 2025b), and trying to make a single task "AI-proof" is what Corbin et al. (2025a) call a "discursive change". This latter only alerts what it is said about the assessment, not what the assessment actually is. "We can make assessment more AI-proof, but if we make them too rigid, we just test compliance rather than

creativity" (Corbin et al., 2026, p.742); the whole detection effort may simply have an expiry date, since once AI tools become embedded and invisible, catching anyone becomes impossible (Eaton. 2023).

This paper argues that chasing AI-proofing has to stop and that redesigning assessment around principles that hold up whether or not students use AI has to start. Researchers have made the case for several of these ideas separately and this study brings these threads together into a single, coherent four-pillar framework, something educators can actually build with rather than a scattering of useful but separate ideas.

The challenge of AI in assessment has been described as a wicked problem, one with no clear definition, no obvious finish line, and no single correct answer, only better and worse ways of coping (Corbin et al., 2026; Rittel & Webber, 1973). There is also a deeper irony in banning AI outright, that is, in much professional work, using whatever tools are available is simply a normal practice, so prohibition can make an assessment less faithful to the job it is meant to prepare students for, not more (Fawns et al., 2025). Wicked problems cannot be solved, only navigated and this study makes no claim otherwise, what it does claim is that this challenge is far easier to navigate when assessment is built around validity, evaluative judgment, authentic process, and a sound program structure than when is built around bans and detection.

2. The Framework: Four Pillars for Assessment beyond AI-Proofing

2.1. Designing around Validity

The first pillar urges a reconsideration of what assessment is for, that is a shift from detecting dishonesty to being genuinely confident that students can do what we say they can. Validity here is not just a technical feature of a test, but it, as Dawson et al. (2024, p.1013) put it: a "property of assessment occasion" and a "social imperative". A piece of work is valid only if it really reflects what the student can do. When seen from this perspective, cheating becomes less a moral failing than a validity problem, as it prevents from honestly certifying that a student has met the standard (Dawson et al., 2024, p.1007) , and this lens folds cheating into something bigger and more positive, that is making sure learning has actually happened (Dawson et al., 2024, p.1011).

This shift puts the spotlight on consequential validity, in other words, the way assessment design shapes how students behave (Dawson et al., 2024, p.1013). Just as high-stakes exams reliably produce last minute cramming, tasks poorly suited to a world full of AI reliability produce AI shortcuts. Students' behavior then, is partly a product of the design not just a personal failing, therefore, getting validity right means going past rules students can ignore and toward structural changes that reshape the task itself, so that the behavior educators hope for comes naturally out of how the task is built (Corbin et al., 2025a, pp. 1089, 1094).

At the program, level this means treating assessment as a chain of evidence rather than a series of one-off tests, or trying to make any single task AI-proof; because confidence is built from many pieces of evidence that fit together (Corbin et al., 2025a, p. 1096). Hence, no single task can put every doubt to rest, but a well-designed sequence of connected tasks can, and the trade-offs between security and authenticity, or between strictness and fairness become much clearer once validity, not cheating prevention, is the aim (Dawson et al,. 2024, p. 1014).

2.2. Evaluative Judgment: The one thing AI cannot do for students

Evaluative judgment is "the capability to make decisions about the quality of work of self and others" (Tai et al., 2018, p.471). It is not a nice-to-have but a cornerstone of what a degree should produce, but the thing that lets graduates keep learning and judging well long after they leave. It has two dimensions: (1) developing an internal sense of what good work looks like, (2) and being able to apply that sense to real work, whether it is their own, a classmate's, or a machine's (Tai et al., 2018: Bearman et al., 2024, p. 4). Students need more than a rubric handed to them, they need the ability to apply such standards on work that is messy, varied, and increasingly shaped by AI (Villarroel et al., 2018, p. 7).

Gen-AI has made this skill more important by widening the gap between how easily we can produce work and how well we can judge it (Bearman et al., 2024, p. 903). as AI tools turn out polished looking material that is also prone to "hallucinations" by producing answers that sound confident but are simply wrong (Bearman et al., 2024, p. 894); Graduates working alongside AI will need deep knowledge of their field to tell a genuine insight from a convincing mistake (Bearman et al., 2024, p. 894). This is exactly why evaluative judgement is something only a human can do, a machine can generate something that looks like quality but it cannot decide what quality is as standards are set by people, shaped by context, and ultimately the practitioner's responsibility (Bearman et al., 2024, pp. 895–896).

This responsibility shapes how tasks are designed, and the final call on whether work is good enough has to sit with the human, not the machine (Bearman et al., 2024, p. 895). When students work with AI systems they cannot see inside, they have to bring their own disciplinary standards to critically assess what the AI gives them (Bearman & Ajjawi, 2023, p. 1168). Therefore,

tasks should build students' ability to weigh up not just AI's output but their own reasoning while using it, because they remain accountable for whatever they submit (Bearman et al., 2024, p. 894; Eaton, 2023, p. 5). Ultimately the real goal, as Bearman and Ajjawi (2023, p. 1168) put it, is not just for students to succeed at a task but to "come to understand how success is constituted" and that understanding stays firmly human.

2.3. Authentic, Process-focused Assessment

The third pillar calls for tasks rooted in real-world work, but with eyes open about what authenticity can and cannot do. Authentic assessment brings realism and genuine problems into the task aiming to mirror professional practice and push students past simply repeating facts toward real mastery (Villarroel et al., 2018, pp. 3, 7). However, in the age of Gen-AI, authenticity has to be treated as something to aim for not a guaranteed fix. As Fawns et al. (2025, p. 397) warn, applying the "authentic" label to a task without carful thoughts can turn it into a "thought-terminating cliché" that shuts down the hard conversations we need to have about trade-offs with grading, fairness, and rigor.

This caution matters because being authentic does not stop AI misuse as in plenty of real jobs, the finished product matters more than how it was made and using whatever tools are on hand including AI is part of doing the work well (Fawns et al., 2025, p. 401). Banning AI can therefore be less true to professional life, not more. Authenticity is best used thoughtfully as a way to open up honest discussion about how tasks relate to real, tool-filled work rather than as a badge, that means, "AI cannot touch this" (Fawns et al., 2025, p. 405).

The way through is to shift attention from the final product to the learning process behind it. Gen-AI has made it easier than ever to produce something that looks good without any real learning (Bearman et al., 2024, p. 903), so assessing the end product alone is no longer sufficient. Tasks should instead require students to make their process visible, and show how they explored a problem, interrogated their sources, and integrated AI-generated material into their own thinking (Perkins et al., 2024, p. 10). Instruments such as e-portfolios, reflective logs, and staged submissions render that process visible and assessable, allowing educators to evaluate a student's digital literacy and judgment alongside the final piece of work (Nieminen et al., 2023, pp. 530, 539). If seen from this angle, process visibility becomes not a form of surveillance but a pedagogical commitment, one that places learning itself rather than its output, exactly at the center of what we assess.

2.4. A Two-Lane Program Structure

The fourth pillar offers a structural response at the level of the whole program, a two-lane design that distributes assessment across two complementary tracks (Corbin et al., 2025a). The first is Secure: meaning assessment that is conducted in person under controlled conditions where AI cannot be used, providing reliable evidence of student capability at critical points in the program. The second is Open: where AI may be used freely or with deliberate scaffolding and the emphasis shifts toward developing learning and working with AI as professionals do (Corbin et al., 2025a, p. 1096). Crucially, both lanes require genuine reconsideration of their design rather than mere relabeling of existing tasks (Corbin et al., 2025a, p. 1096).

This rests on programmatic assessment theory, which holds that sound assessment depends on "a deliberate and arranged set of longitudinal assessment activities" rather than a handful of isolated high-stakes tests (van der Vleuten et al., 2012, p. 205). Individual tasks function as data points that are rich for learning in the moment, while the accumulation of evidence across them underpins the higher-stakes decisions about certification (van der Vleuten et al., 2012). It is the coherent structure across the whole program, not any single task that renders the system fit for purpose (van der Vleuten et al., 2012, p. 212).

This layered logic aligns with what Dawson et al. (2024, p. 1012) describe as the "Swiss Cheese" model of academic integrity, meaning any single safeguard has gaps but multiple layers working together compensate for one another's weaknesses. This is a considerably more defensible position than attempting to make any one task AI-proof, an effort that yields either rules no one can enforce, or tasks so constrained that they compromise validity in other respects. Treating validity as a chain of evidence across the program (Corbin et al., 2025a, p. 1096) also allows later tasks to build on earlier ones, making it progressively more difficult for AI-generated work to substitute for genuine, long-term learning. The integrity of the system as a whole then derives from how it is constructed, not from how strictly its prohibitions are enforced.

3. Discussion

These four pillars are not a menu from which to choose. They operate together and each is working at a different level and reinforcing the others. Validity is the backbone that defines what assessment is for; and reframing Gen-AI as a problem of design rather than of student behavior. Evaluative judgment is the capability students must develop, that is the distinctly human ability to judge quality that cannot be delegated to a machine. Authentic, process-focused assessment describes the form the tasks themselves should take, from being grounded in real-world practice to being attentive to the learning journey, and clear-

eyed about the limits of authenticity. The two-lane structure is the container that holds the others together by distributing assessment across secure and open modes and gathering evidence across multiple points.

Taken together, the pillars change the governing question from "did the student cheat?" to "can the student demonstrate genuine capability?" The first question casts the student as a suspect and the educator as an enforcer producing the illusion of control that Corbin et al. (2025a) describe. The second casts the student as a learner and the educator as a designer, locating integrity in the structure of the tasks rather than in students' willingness to comply. The framework speaks directly to Corbin et al.'s (2026) finding that AI in assessment is a wicked problem, one that places an unavoidable burden of judgment on teachers. Building assessment around these four pillars does not lift that burden, nothing can. However, it gives educators something solid to work with: permission in Corbin et al.'s (2026) terms, to compromise, to diverge by context, and to keep adjusting, all of which becomes far more achievable when validity, judgment, authentic process, and sound structure replace detection and prohibition as the priorities.

The study acknowledges a number of limitations. First, almost all the evidence is drawn from higher education, and predominantly from Australia and the United Kingdom, so its application to schools, vocational training, or non-Western settings warrants caution. Second, the pillars are derived from theory; there is as yet limited empirical evidence that programs designed around all four assure learning more effectively and, given the wicked nature of the problem, such evidence will always be partial. Third, the framework draws heavily on a relatively concentrated group of researchers at Deakin University's CRADLE center and the University of Sydney. This reflects where much of the leading work is being conducted, but the ideas would be strengthened by engaging a wider range of voices, particularly from outside the English-speaking world.

The research agenda that follows is clear. Studies are needed that test these pillars within particular disciplines. Instruments are needed that can measure whether students' evaluative judgment actually develops in AI-rich environments. The application of two-lane program structures to schools and vocational training warrants investigation. In addition, the way validity plays out across different cultures including the place of AI literacy within different educational traditions deserves sustained attention (Almatrafi et al., 2024).

4. Conclusion

Higher education's dominant response to Gen-AI prohibition, detection, and policy rules shares a single flaw: it depends on students obeying instructions that no one is able to enforce, this makes it not merely ineffective but counterproductive, producing an illusion of control while consuming energy that could be directed toward genuine redesign.

The "Beyond AI-Proofing" framework offers a more constructive alternative, built on four pillars. Designing around validity reorients assessment toward confidence in learning rather than the detection of misconduct. Evaluative judgment establishes the capacity to judge quality, one's own, others', and AI's as something central that we teach. Authentic, process-focused assessment builds tasks that follow the learning journey and treats authenticity as a goal involving trade-offs rather than as a shield against AI. Moreover, a two-lane program structure secures evidence of learning at the program level while leaving room for open, supported engagement with AI.

What the four pillars share is that they build trustworthiness into the design of assessment, rather than attempting to impose it through rules that cannot be kept. A system constructed in this way does not stand or fall on whether students use AI. It stands or falls on whether students can demonstrate genuine capability across a connected, long-term chain of evidence, a question that remains answerable whatever tools exist. What this requires depends on one's role. For educators, it means treating validity as the criterion that guides design, and accepting that a degree of messiness, openness, and ongoing revision are not failures but features of a system built for reality. For program designers, it means thinking across the whole degree, about how individual tasks combine to assure learning. For policymakers, it means moving away from frameworks that depend on enforcement and investing instead in redesign genuinely suited to a world with AI in it. For researchers, it means testing these ideas and extending them beyond their origins. Assessment was never really about catching students doing the wrong thing. It was always about being confident that the people we graduate can do what we say they can, a purpose that Gen-AI has made far harder to ignore, and far harder to evade.

Funding: This research received no external funding.
Conflicts of Interest: The author declares no conflict of interest.
ORCID iD: https://orcid.org/0009-0008-1839-7014

References

[1] Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open*, *6*, 100173. https://doi.org/10.1016/j.caeo.2024.100173

[2] Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

[3] Bearman, M., Nieminen, J. H., & Ajjawi, R. (2023). Designing assessment in a digital world: An organising framework. *Assessment & Evaluation in Higher Education*, *48*(3), 291–304. https://doi.org/10.1080/02602938.2022.2069674

[4] Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

[5] Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2026). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, *51*(4), 736–752. https://doi.org/10.1080/02602938.2025.2553340

[6] Corbin, T., Dawson, P., & Liu, D. (2025a). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, *50*(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

[7] Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025b). "Where's the line? It's an absurd line": Towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education*, *50*(5), 705–717. https://doi.org/10.1080/02602938.2025.2456207

[8] Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, *49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

[9] Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, *19*, Article 23. https://doi.org/10.1007/s40979-023-00144-1

[10] Fawns, T., Bearman, M., Dawson, P., Nieminen, J. H., Ashford-Rowe, K., Willey, K., Jensen, L. X., Damşa, C., & Press, N. (2025). Authentic assessment: From panacea to criticality. *Assessment & Evaluation in Higher Education*, *50*(3), 396–408. https://doi.org/10.1080/02602938.2024.2404634

[11] Nieminen, J. H., Bearman, M., & Ajjawi, R. (2023). Designing the digital in authentic assessment: Is it fit for purpose? *Assessment & Evaluation in Higher Education*, *48*(4), 529–543. https://doi.org/10.1080/02602938.2022.2089627

[12] Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, *21*(6). https://doi.org/10.53761/q3azde36

[13] Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

[14] Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

[15] Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

[16] Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396