ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Teaching beyond AI: Gamified problem-based formative assessment aligned with UNESCO AI competency frameworks

Usman Durrani
Global Colleges Australia

Mustafa Akpinar
Higher Colleges Technology UAE

Ghazi Al Naymat
Ajman University UAE

The arrival of capable generative AI (GenAI) tools has posed a genuine and largely unresolved challenge for summative assessment in higher education. When AI-generated submissions evade institutional detection 94% of the time and examination scores rise sharply without any change in teaching, the evidentiary basis of take-home and invigilated assessment is genuinely in question. This paper draws on two lines of work to argue that classroom-based gamified problem-based formative assessment (GPBFA) offers a more durable response than detection-focused approaches. The first line is an empirical study of saval.app, a GPBFA platform evaluated through Keller's ARCS motivational model across a full university semester (pre n = 41, post n = 44). Results across all four ARCS dimensions were encouraging (all $p < .0001$; Cronbach's alpha > .93; composite 81% positive response rate), with near-zero motivational decline across 12 weeks. The second line is a review of UNESCO's 2024 AI Competency Frameworks for Teachers and Students, which we argue are philosophically and pedagogically aligned with classroom GPBFA in ways the existing literature has not yet made explicit. Taken together, we suggest that moving assessment into the classroom — where it is live, observed, and problem-centred — may address both the integrity challenge and the AI literacy imperative that UNESCO frameworks define.

Keywords: generative AI, academic integrity, formative assessment, gamification, problem-based learning, ARCS model, saval.app, UNESCO AI competency framework, higher education

Introduction

Assessment in higher education is facing a challenge it was not designed for. Summative tasks — essays, take-home assignments, online quizzes — were built on the assumption that the work submitted reflects the knowledge and effort of the student who submitted it. That assumption has always been imperfect; now it is under real pressure. GPT-4 scores at the 90th percentile on the Uniform Bar Exam. A longitudinal study tracking 3,109 students found examination scores rose by nearly 22 percentage points after ChatGPT's release, with course content and teaching staff unchanged throughout (Dunn, 2025). A University of Reading study found that AI-generated submissions evade institutional detection processes 94% of the time (Scarfe et al., 2024). Australia's TEQSA, in September 2025 guidance, reached a conclusion many in the sector had been reluctant to state plainly: AI-assisted cheating has become very difficult to detect consistently, and the sector needs to rethink assessment rather than sharpen its detection tools (Lodge et al., 2025).

At roughly the same time, UNESCO released a pair of frameworks that matter for how this challenge is understood. The AI Competency Framework for Teachers (AI CFT) and the AI Competency Framework for Students (AI CFS), both published in August 2024, represent the first coordinated international effort to define what AI literacy should look like across all levels of education (UNESCO, 2024a, 2024b). What is notable about these frameworks is that they are not primarily about using AI — they are about maintaining human agency and ethical judgment in an environment where AI is increasingly available, and about keeping authentic intellectual engagement at the centre of learning. That is, the capacities UNESCO most wants educators and

students to develop are the same ones that a classroom-based, problem-centred assessment model is designed to build and observe.

This paper connects these two concerns. We begin with a brief account of why summative assessment is under pressure and why detection-based remedies are unlikely to be sufficient. We then present an empirical study of saval.app, a GPBFA platform we deployed across a full semester with an undergraduate cohort, evaluated using the ARCS motivational model. We follow this with a review of the UNESCO framework literature and an argument about how classroom GPBFA maps onto the competency levels those frameworks define. The paper closes with a discussion of what this means for curriculum redesign in the Australian context and where the research needs to go next.

The summative assessment problem under generative AI

Why detection is not the answer

Institutional responses to AI-assisted academic misconduct have centred on detection — Turnitin's AI detection feature, originality scores, disclosure policies requiring students to declare AI use. The evidence that this approach can hold up over time is not encouraging. Detection tools tested against genuinely human-authored text have reported false positive rates above 60% in some studies (Weber-Wulff et al., 2023), while the same tools miss AI-generated content at high rates when that content has been lightly paraphrased or processed through intermediary tools. Australian Catholic University's experience illustrates the problem: nearly 6,000 students were flagged for AI misconduct in 2024, most cases were subsequently overturned; and Curtin University withdrew AI detection tools entirely on accuracy grounds (Bergin, 2025). Corbin et al. (2025) put the underlying issue clearly: the problem is not primarily a detection problem. It is a design problem. As long as the form of assessment can be completed by AI, the integrity of that assessment is in question regardless of what detection measures are applied.

TEQSA's September 2025 guidance acknowledged this directly, identifying three reform pathways: program-wide redesign, unit-level assurance measures, and hybrid models combining in-person and remote components (Lodge et al., 2025). All three involve changing what assessment looks like, not improving how misconduct is identified after the fact.

The longer-standing pedagogical case

It is worth noting that GenAI did not create the pedagogical problems with high-stakes summative assessment — it exposed them. Terminal examinations and take-home assignments have long been associated with surface learning strategies and performance anxiety — cramming for reproduction rather than genuine understanding — patterns that self-determination theory attributes to the extrinsically regulated conditions these tasks produce (Deci & Ryan, 2000). The research on formative assessment, by contrast, has generated some of the more consistent findings in educational research. Black and Wiliam's (1998) synthesis of over 250 studies reported effect sizes in the range of 0.40 to 0.70 standard deviations for formative approaches — gains that, in Hattie's (2008) terms, represent meaningful improvements rather than marginal ones. The argument for moving assessment into the classroom, closer to the moment of learning, is not new. What GenAI has done is make that argument harder to set aside.

Classroom-based GPBFA: why the design matters

The integrity case

There is something straightforward about the integrity argument for classroom-based assessment that sometimes gets lost in the literature's focus on detection. If a student is completing an assessment task in a room, in real time, with an instructor present, there is no practical mechanism by which that task can be outsourced to an AI. This is not a better detection method — it is a different model of assessment entirely. Classroom GPBFA does not try to identify AI use after the fact; it creates conditions in which AI use during the assessment event is not possible. The assessment moment and the learning moment are the same observed event. This matters for what the submitted work can be taken as evidence of.

Problem-based learning and gamification

Research on problem-based and gamified approaches in higher education offers a reasonably well-established foundation. PBL, in Dochy et al.'s (2003) meta-analysis, produced consistent benefits for content retention and collaborative reasoning, with the most pronounced gains for higher-order work — applying knowledge in unfamiliar situations, analysing cases with no single right answer. Walker and Leary (2009) found similar patterns across a range of disciplines. Gamification contributes a motivational dimension. Drawing on Malone's (1981) early theoretical work on intrinsic engagement, challenge-based mechanics increase willingness to attempt difficult problems and to persist through initial failure — a disposition that is both educationally valuable and, notably, difficult to fake. Neither PBL nor gamification works at a distance or in isolation; both depend on a live, social context that gives the activity its meaning.

AI changes the feasibility calculation here in an interesting way. When deployed on the educator's side — generating varied problem scenarios, grading open-ended responses across a class, surfacing real-time analytics — AI makes GPBFA more scalable than it would otherwise be. The instructor is not individually reading 30 responses to the same question; the AI handles the initial assessment and flags patterns for human attention. The live classroom element is preserved, and AI amplifies rather than undermines it (Zhai & Nehm, 2023).

UNESCO AI competency frameworks

What the frameworks say

Released in August 2024, the two UNESCO frameworks are worth examining carefully, because they carry a philosophical stance that distinguishes them from narrower AI literacy curricula. The AI CFT for teachers defines 15 competencies across five dimensions: Human-Centred Mindset, Ethics of AI, AI Foundations and Applications, AI Pedagogy, and AI for Professional Learning. Each competency is organised across three developmental levels — Acquire, Deepen, and Create — that move from conceptual familiarity to independent, creative application (UNESCO, 2024b). The AI CFS for students is structured around four dimensions — Human-Centred Mindset, Ethics of AI, AI Techniques and Applications, and AI System Design — across parallel levels of Understand, Apply, and Create (UNESCO, 2024a).

What matters about the underlying philosophy is that UNESCO treats human agency and ethical judgment as central to AI competency rather than as aspirational additions. Designed amid growing concern that AI in education could reduce students to passive consumers of AI-generated content, the frameworks push back by centring capacities — critical thinking, reflective self-assessment — that AI cannot perform on a student's behalf.

Table 1
UNESCO AI competency framework dimensions and progression levels

Framework	Dimensions	Progression levels
AI CFT (Teachers)	Human-Centred Mindset; Ethics of AI; AI Foundations; AI Pedagogy; AI for Professional Learning	Acquire → Deepen → Create
AI CFS (Students)	Human-Centred Mindset; Ethics of AI; AI Techniques and Applications; AI System Design	Understand → Apply → Create

What the emerging research shows

Empirical engagement with these frameworks is still relatively thin, which partly reflects how recently they were published. Okada et al. (2025) offer the most methodologically developed study to date: a mixed-methods examination of 330 secondary students across the UK, Greece, and Brazil, finding that AI competency development tended to occur through social learning contexts rather than through formal instruction. That

finding resonates with the classroom GPBFA model. Wang et al. (2026) looked at AI integration across Chinese higher education institutions and found a pattern worth noting: implementations were concentrated in technical infrastructure, with considerably weaker engagement with the Human-Centred Mindset and Ethics dimensions. Chiu et al. (2025) developed and validated a self-efficacy scale for teachers (*n* = 434) explicitly aligned with the UNESCO teacher framework, finding it necessary to add assessment as a sixth domain — a gap we take as further support for the GPBFA argument.

Mapping these levels onto classroom GPBFA is not difficult. Students building conceptual familiarity with AI — the UNESCO Understand level — find a reasonable parallel in traditional instruction. Where classroom GPBFA is most directly relevant, however, is at the Apply level, which demands active judgment in realistic contexts: it produces exactly that kind of observed engagement. As for the Create level — independent capability and reflective transfer — this is what the saval.app study's Confidence findings speak to most directly.

Saval.app: platform design and features

Saval.app is a web-based formative assessment AI platform designed specifically for live classroom use. The design decision to make it classroom-only — rather than a homework or asynchronous tool — is deliberate and reflects the integrity argument above. Operationally, the instructor selects a challenge format, difficulty setting, and grading strictness before each scheduled tutorial, then launches a session-specific quiz code that students join from their own devices; the AI grading engine scores responses and updates the leaderboard as the session runs, so the activity is embedded directly within, rather than added alongside, timetabled class time. In this study, saval.app sessions were run this way across 12 weekly 90-minute tutorials, using scenarios drawn directly from the unit's own content (stakeholder mapping, requirements documentation, Python data tasks, and project case problems); the full instrumentation and analysis procedure is set out in Research methodology below. The platform supports multiple gamified assessment modes. Class Quiz runs as a live per-student competition with a visible leaderboard. Fly or Fall presents student avatars that rise on correct responses and fall on incorrect ones, adding a survival dynamic that students tend to find highly engaging. Group Bucket shifts the competition to the team level with ranked dashboards. The Great Wall introduces inter-team challenge mechanics, where teams build points collectively and can disrupt rival teams, adding a strategic dimension.

Six challenge formats are available: open scenario problems, multiple choice, true/false, match-the-following, video-based scenarios, and image-based problems. This range makes the platform usable across most undergraduate disciplines without significant content adaptation. The AI grading engine assesses text responses, source code, mathematical expressions, and image submissions, returning criterion-referenced feedback within a few seconds. Instructors can configure grading strictness (Flexible, Normal, or Strict) and write discipline-specific rubrics that the AI applies. Live dashboards update every 10 seconds. Post-session analytics include heatmaps by challenge and student performance segmentation, alongside a natural-language query assistant that lets instructors interrogate session data in plain terms. Integrity features include copy/paste prevention and session-specific quiz codes, with full audit trails queryable after each session.

Theoretical framework: the ARCS model

We evaluated saval.app's motivational outcomes using Keller's (1987) ARCS model, which identifies four conditions that need to be in place for learners to remain motivated over time: Attention (genuine curiosity and interest in the activity), Relevance (a perceived connection to things the learner cares about), Confidence (a reasonable sense that success is possible and that effort makes a difference), and Satisfaction (rewards — intrinsic or extrinsic — that make continued engagement feel worthwhile). The ARCS model was chosen partly because it is well-validated in technology-mediated learning contexts and partly because its four dimensions map in a reasonably transparent way onto saval.app's design features. Attention is addressed through the gamification mechanics, which are genuinely novel for most cohorts and maintain a competitive social dynamic. Relevance depends on the quality and authenticity of the problem scenarios used — this is partly a platform feature and partly an instructor responsibility. Confidence is supported by the immediacy of the AI feedback and by challenge structures that give students early opportunities to succeed before difficulty

increases. Satisfaction comes through visible outcomes — leaderboards, team scores — and through the social recognition that visible performance creates in a live classroom setting.

Research methodology

The effectiveness of the intervention was evaluated through a pre-post survey design measuring change across Keller's four ARCS motivational dimensions, triangulated with thematic analysis of open-ended student reflections collected in the post-unit survey. This combination was chosen to capture both the magnitude and stability of motivational change over the semester (via the quantitative ARCS measures) and the qualitative texture of how students experienced the intervention (via thematic analysis), rather than relying on a single end-of-semester satisfaction rating. The design, instruments, and analysis procedures are set out in full below.

Design, context, and participants

Our research used a pre-post survey design with a single cohort of undergraduate students enrolled in a Business Analysis unit at a metropolitan Australian university. The unit covered Python-based data analysis and business scenario work — a combination that creates genuine variation in student confidence and that covers content where AI tools perform reasonably well, making the formative assessment context all the more pointed. Forty-one students completed the pre-unit survey in Week 1, before any exposure to saval.app. Forty-four completed the post-unit survey in Week 12. The cohort was drawn from second- and third-year business and information technology programmes. Ethics approval was obtained from the university's human research ethics committee prior to data collection, and participation was voluntary. Students provided written informed consent before completing either survey.

Instruments, deployment, and analysis

A pre-unit survey comprised 29 five-point Likert items (Strongly Disagree to Strongly Agree) distributed across the four ARCS dimensions: seven items each for Attention and Relevance, eight for Confidence, and seven for Satisfaction. The post-unit survey used 20 items (five per dimension) plus three open-ended questions asking students to reflect on their experience with saval.app. Saval.app sessions ran across twelve weekly 90-minute tutorials, using scenarios drawn from business analysis content — stakeholder mapping, requirements documentation, Python data tasks, and project case problems. Descriptive statistics and Cronbach's alpha were computed for each dimension. One-sample t-tests compared post-unit dimension means against the neutral midpoint of 3.0. Qualitative responses were analysed following Braun and Clarke's (2006) six-phase thematic analysis procedure, with two cycles of coding.

Findings

Instrument reliability

Internal consistency was high across both administrations. Pre-survey Cronbach's alpha values were: Attention .934, Relevance .958, Confidence .964, and Satisfaction .966. Post-survey values were: Attention .982, Relevance .946, Confidence .931, and Satisfaction .963. All values exceed the widely used threshold of .70 (Nunnally & Bernstein, 1994), and the consistency of the pattern across both time points supports confident interpretation of the dimension-level findings.

Pre-unit baseline (*n* = 41)

Students entered the unit with notably positive motivational dispositions across all four dimensions (composite M = 4.09, SD = 0.69). The lowest individual item ratings were both Python-related, which reflects a well-documented pattern of coding anxiety among business students. Even so, the composite baseline was well above neutral, which provides a useful reference point for interpreting the post-unit results.

Table 2
Pre-unit ARCS descriptives (n = 41)

ARCS dimension	Items	M	SD	% Positive	Note
Attention	7	4.14	0.58	84%	High anticipatory curiosity
Relevance	7	4.11	0.61	83%	Strong perceived course value
Confidence	8	4.08	0.63	80%	Python items lowest-rated
Satisfaction	7	4.04	0.64	79%	Positive outcome expectations
Composite	29	4.09	0.69	81%	

Post-unit results (*n* = 44)

All four dimension means in the post-unit survey significantly exceeded the neutral midpoint of 3.0, with all t-tests returning $p < .0001$. The single highest-rated item — "I feel confident applying similar concepts independently after using saval.app" — reached an 89% positive response rate. The item measuring perceived course alignment achieved 86%. These figures are worth pausing on: the confidence transfer item speaks directly to the question of whether students are developing independent capability, as distinct from becoming dependent on the platform or the AI grading engine.

Table 3
Post-unit ARCS statistics and one-sample t-tests against neutral midpoint 3.0 (n = 44)

Dimension	M	SD	% Positive	t(43)	p	Highest-rated item
Attention	3.91	0.89	78.6%	6.84	< .0001	Live session engagement
Relevance	4.03	0.63	82.3%	10.82	< .0001	Course objectives alignment (86%)
Confidence	4.07	0.61	81.8%	11.65	< .0001	Independent transfer (89%)
Satisfaction	4.03	0.66	81.8%	10.42	< .0001	Collaborative task enjoyment

Pre-post comparison and qualitative findings

Pre-post changes across the four dimensions are set out in Table 4. The Attention dimension showed the most notable decline ($\Delta = -0.23$), which is consistent with what Clark (1994) documented as novelty attenuation in educational technology: initial excitement moderates as the tool becomes familiar. The post-survey mean of 3.91 remained significantly above neutral, however, so this is better understood as a settling into routine engagement than as disengagement. The more striking finding is the near-zero change in both Confidence ($\Delta = -0.01$) and Satisfaction ($\Delta = -0.01$). Students who entered the unit expecting to feel capable through this kind of active learning found, 12 weeks later, that those expectations had held. That durability is arguably more meaningful than a strong initial response would have been.

Table 4
Pre-post ARCS dimension comparison

Dimension	Pre M (SD)	Post M (SD)	Δ M	Post % pos.
Attention	4.14 (.58)	3.91 (.89)	−0.23	79%
Relevance	4.11 (.61)	4.03 (.63)	−0.07	82%
Confidence	4.08 (.63)	4.07 (.61)	−0.01	82%
Satisfaction	4.04 (.64)	4.03 (.66)	−0.01	82%
Composite	4.09 (.69)	4.01 (.75)	−0.08	81%

Thematic analysis of the open-ended responses identified four patterns in how students described their experience. The first, which we labelled active engagement, captured the sense that sessions required students to do something rather than observe something — a distinction they drew repeatedly and often with some enthusiasm. The second, understanding through application, reflected students articulating a difference between applying a concept in a scenario and memorising it for later — a distinction that is not trivial and that maps onto what higher-order learning is supposed to feel like. The third theme centred on visibility: the leaderboard and team elements made achievement socially visible in a way that private grades do not, and students found this motivating rather than threatening, though the study did not examine whether this held equally for all students. The fourth theme was what we might call openness to more: students expressed interest in having similar sessions in other units, which suggests some transfer of motivational orientation beyond this particular course.

Discussion

What the findings suggest

Taken as a whole, these results are encouraging, though the usual caveats about single-cohort designs apply. What stands out most is not the overall positive response — that could reasonably be attributed to novelty — but the specific pattern of the pre-post comparison. Novelty effects tend to show up in Attention, and that is where the study's most pronounced decline occurred. The fact that Confidence and Satisfaction held almost perfectly across 12 weeks suggests these dimensions were tracking something more durable than initial excitement. Students' sense that they could apply what they had learned, and that the experience was worth having, remained stable across a semester of regular exposure. That is a different kind of evidence than a strong end-of-semester rating.

Deserving specific attention in the context of the GenAI problem is the 89% positive response on the independent transfer item. If the concern with AI-assisted assessment is that students are producing work that does not reflect their own understanding, then the relevant counter-evidence is not a high grade but a credible indication that students can apply what they have supposedly learned. An 89% rate of students reporting confidence in independent application, measured after 12 weeks of classroom engagement, is at least a meaningful data point in that direction — though we are careful not to overstate what a self-report measure can demonstrate.

The UNESCO framework connection

As noted earlier, UNESCO defines three levels of competency development for students: Understand, Apply, and Create. The classroom GPBFA model, as implemented through saval.app, maps most naturally onto the Apply level — students are exercising judgment in real-time, contextualised problem scenarios rather than acquiring or reproducing knowledge. The qualitative theme of understanding through application supports this reading. The 89% confidence transfer finding speaks to the Create level's emphasis on independent capability: students are not just performing tasks with platform support, but are beginning to develop the kind of transferable self-efficacy that the Create level describes.

What the UNESCO framework literature has not yet made clear is how these competency levels should actually be cultivated in practice. The frameworks are aspirational instruments: they describe target states but are largely silent on pedagogy. Classroom GPBFA, as one evidence-based approach, offers at least a partial answer: the Apply level develops through repeated, authentic problem engagement with immediate feedback; the Create level develops through the accumulation of successful performance experiences that build genuine self-efficacy over time. The ethics dimension of both frameworks — which emphasises transparency, human oversight, and responsible AI use — finds a practical correlate in saval.app's instructor-configurable AI grading and audit trail features, which keep a human educator's judgment in the loop rather than delegating assessment to an opaque automated system.

Limitations that matter

Several limitations of this study are worth being direct about. The single-cohort design means we cannot attribute motivational outcomes specifically to saval.app rather than to the instructor's approach, the cohort's characteristics, or the broader PBL design. There was no control condition, so we cannot compare GPBFA outcomes to a conventionally assessed version of the same unit. The study measures motivation, not learning achievement or academic performance — and while motivation and learning are related, they are not the same thing. The most productive next step would be to integrate saval.app's session-level analytics with academic performance data, to examine whether the motivational gains observed here correspond to measurable differences in what students can actually do. A multi-cohort or multi-institution design would also substantially strengthen the evidence base. It is worth being explicit about what this study can and cannot claim as its educational contribution. The ARCS findings are, on their own, evidence of student perceptions rather than of learning gains. The anticipated contribution beyond those perceptions lies in the link we draw, in the UNESCO framework connection above, between the near-zero decline in Confidence, the 89% independent-transfer response, and the Apply and Create competency levels those frameworks define: we treat these as early behavioural indicators consistent with developing independent-application competency, not as proof of it. Confirming that link — connecting motivational stability of this kind to demonstrated performance or transfer — is precisely the empirical gap the proposed integration with academic performance data is intended to close.

Conclusion

Generative AI's pressure on summative assessment is, in our view, best understood as an opportunity to take formative assessment more seriously than Australian higher education has typically done. The evidence that detection-based approaches are unreliable is now fairly strong. The evidence that formative assessment, well-designed, produces meaningful learning outcomes is also fairly strong, and has been for some time. What classroom-based GPBFA adds to this picture is a model in which the integrity question and the pedagogical question have the same answer: bring assessment into the classroom, make it live and problem-centred, and let students demonstrate understanding in a context where the work is visibly and verifiably their own.
Our saval.app study offers a modest but coherent piece of evidence in support of that model. Motivational outcomes were positive and stable across a full semester, with the dimension most directly relevant to genuine learning — Confidence in independent application — showing almost no decline over 12 weeks. The UNESCO framework review suggests that this kind of assessment is also well-aligned with the AI competency goals that global education policy is now organised around. For curriculum designers and academic leaders in the Australian context — working under TEQSA's 2025 mandate to rethink assessment at the unit and programme level — classroom GPBFA is worth considering not as a niche innovation but as a practically available, evidence-supported response to a challenge the sector needs to take seriously.

References

Bergin, J. (2025, October 9). *University wrongly accuses students of using artificial intelligence to cheat*. ABC News. https://www.abc.net.au/news/2025-10-09/artificial-intelligence-cheating-australian-catholic-university/105863524

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Chiu, T. K., Ahmad, Z., & Çoban, M. (2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, *30*(5), 6667-6685. https://doi.org/10.1007/s10639-024-13094-z

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development*, *42*(2), 21-29. https://doi.org/10.1007/BF02299088

Corbin, T., Dawson, P., & Liu, D. (2025). 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

Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, *13*(5), 533-568. https://doi.org/10.1016/S0959-4752(02)00025-7

Dunn, P. (2025). College students’ test scores soared after ChatGPT launch. *International Journal of Mathematics Education in Science and Technology*, *57*(8), 1732-1745, DOI: 10.1080/0020739X.2025.2539711

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, 10(3), 2–10. https://doi.org/10.1007/BF02905780

Lodge, J. M., Howard, S. K., Bearman, M., & Dawson, P. (2025). *Enacting assessment reform in a time of artificial intelligence*. TEQSA, Australian Government. https://www.teqsa.gov.au/sites/default/files/2025-09/enacting-assessment-reform-in-a-time-of-artificial-intelligence.pdf

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333-369. https://doi.org/10.1207/s15516709cog0504_2

Mikeladze, T., Meijer, P. C., & Verhoeff, R. P. (2024). A comprehensive exploration of artificial intelligence competence frameworks for educators: A critical review. *European Journal of Education*, *59*(3), e12663. https://doi.org/10.1111/ejed.12663

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

Okada, A., Sherborne, T., Panselinas, G., & Kolionis, G. (2025). Fostering transversal skills through open schooling supported by the CARE-KNOW-DO pedagogical model and the UNESCO AI competencies framework. *International Journal of Artificial Intelligence in Education*, 1-46. https://doi.org/10.1007/s40593-025-00458-w

SAVAL.app. (2024). *Platform description for academic reference*. SAVAL Education Inc.

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024) A real-world test of artificial intelligence infiltration of a university examinations system: A “Turing Test” case study. *PLoS ONE 19*(6): e0305354. https://doi.org/10.1371/journal.pone.0305354

UNESCO. (2024a). AI competency framework for students. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000391105

UNESCO. (2024b). AI competency framework for teachers. UNESCO. https://www.unesco.org/en/articles/ai-competency-framework-teachers

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *Interdisciplinary Journal of Problem-based Learning*, *3*(1), 6. https://doi.org/10.7771/1541-5015.1061

Wang, J., Hong, W. C. H., Xu, X., Zhang, Y., & Yin, Y. (2026). Exploring AI competency in Chinese undergraduates through the UNESCO framework. *Acta Psychologica*, *263*, 106299. https://doi.org/10.1016/j.actpsy.2026.106299

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-39. https://doi.org/10.1007/s40979-023-00146-z

Wiliam, D., & Leahy, S. (2024). *Embedding formative assessment: Practical techniques for K–12 classrooms (Practical formative assessment techniques for K–12 classrooms)*. Solution Tree Press.

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390-1398. https://doi.org/10.1002/tea.21885

ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Durrani, U., Akpinar, M. & Al Naymat, G. (2026). Teaching beyond AI: Gamified problem-based formative assessment aligned with UNESCO AI competency frameworks. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 916–925). https://doi.org/10.65106/apubs.2026.3202

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Durrani, U., Akpinar, M. & Al Naymat, G 2026