Vol.9 No.2 (2026)

Journal of Applied Learning & Teaching

ISSN: 2591-801X

Proudly owned and sponsored by Kaplan Business School, Australia

Content Available at: https://jalt.open-publishing.org/index.php/jalt/index

Assessment twins: An approach for strengthening assessment validity in the age of generative AI

Jasper Roe [A]
Mike Perkins [B]
Louie Giray [C]

[A] School of Education, Durham University, United Kingdom
[B] Centre for Research & Innovation, British University Vietnam, Vietnam
[C] Mapúa University, Philippines & Khazar University, Azerbaijan

Keywords

Academic integrity;
AI pedagogy;
artificial intelligence;
assessment design;
assessment validity;
generative AI;
higher education.

Correspondence

jasper.j.roe@durham.ac.uk [A]

Article Info

Received 5 February 2026
Received in revised form 23 March 2026
Accepted 24 March 2026
Available online 3 June 2026

DOI: https://doi.org/10.37074/jalt.2026.9.2.3

Abstract

The rise of generative artificial intelligence (GenAI) is raising pressing concerns about the integrity and validity of higher education assessment. Assessment redesign is increasingly seen as necessary; however, there is a relative lack of literature detailing practical approaches. In this study, we introduce the concept of assessment twins as a practical approach to redesigning assessment tasks. We use Messick's unified validity framework to systematically map the ways in which GenAI threatens content, structural, consequential, generalisability, substantive, and external validity. Following this, we conceptualise assessment twins as two deliberately linked components that address the same learning outcomes through different modes of evidence, scheduled closely together to allow for cross-verification. We explain how the twin approach helps mitigate validity threats by triangulating evidence across pedagogically valuable, yet GenAI-vulnerable, assessment formats. To guide implementation, we propose an assessment twin design process: identifying vulnerabilities, aligning outcomes, selecting complementary tasks, and developing interdependent marking schemes. We also acknowledge the challenges, including resource intensity, equity concerns, and the need for empirical validation. Nonetheless, we contend that assessment twins represent a validity-focused response to GenAI that prioritises pedagogy while supporting meaningful student learning outcomes.

Introduction

Generative artificial intelligence (GenAI) remains a burgeoning area of research in assessment practice in higher education. Since the public release of advanced GenAI technologies, concerns regarding academic integrity and the assurance of learning have come to the fore, as learners can now freely access tools that produce human-like, stylistically and grammatically sophisticated text (Giray, 2024a; Perkins, 2023). Given that many forms of higher education assessment have traditionally relied on formats such as the take-home essay, written research project, or prepared presentation, there is a high risk that GenAI could be used to facilitate academic integrity violations, such as the misrepresentation of authorship (Roe, 2025).

The paradigm-rupturing nature of GenAI has also prompted deeper reflection on the purpose of assessment and the role of higher education itself (Bannister et al., 2025; Giray et al., 2024), although there are few clear areas of consensus on what this technology means for learning and assessment. Critics argue that reliance on GenAI may lead to reduced learner agency (Roe & Perkins, 2024), dependency, and the erosion of critical thinking skills (Giray, 2025; Gonsalves, 2024). In contrast, national governments (for example, the United Kingdom) have set a policy direction that embraces GenAI, arguing that its usage can transform education and boost learner outcomes, while also relieving pressure on educators. This includes assessment-related activities, such as providing feedback (Department for Education [DFE], 2025).

The long-term effects and direction of travel regarding GenAI in teaching, learning, and assessment are unclear, and there are no silver bullet solutions. However, assessment design requires urgent attention to avoid sliding into performative or uncritical AI integration, in which AI use is tolerated or applied to existing tasks without consideration of the implications for learning (Rudolph et al., 2025). Amidst this messy and shifting landscape, we argue that there is a need for practical solutions that can form part of a multi-strand approach to ensuring assessment validity in the age of AI.

In this study, we propose the concept of assessment twins for GenAI-vulnerable tasks. This approach is developed from but is distinct from traditional protocols common in higher education assessments (such as the oral viva voce, that is, an oral defence or examination in which a student verbally defends their work). The reason that this is distinct is related to the core organising logic: assessment twins are fundamentally designed as a validity-driven response to a GenAI vulnerability from the outset, and this underpins the entire set of assessment tasks. A twin approach entails pairing each task that may be susceptible to GenAI assistance or completion (such as a take-home essay) with a second, less vulnerable task that assesses the same outcomes. This design has several advantages. First, it strengthens assessment validity by generating confirmatory data on the same set of learning outcomes. Second, it preserves the pedagogical value of established assessment formats that should not be discarded entirely and enables judicious, authentic, and appropriate engagement with GenAI, which has been named as a core principle of assessment redesign in the AI era (Tertiary Education Quality and Standards Agency [TEQSA], 2025). Third, it aligns with broader calls for educators to emphasise collaborative, in-person, and multimodal forms of assessment in the age of GenAI (Rudolph et al., 2023) and the use of multiple, inclusive, and contextualised methods of assessment to form “trustworthy judgements about student learning” (TEQSA, 2025, p. 1).

The remainder of this paper is organised as follows. We begin by reviewing the current literature on GenAI and assessment validity. We then introduce the concept of an assessment twin in depth and explain how it enhances multiple strands of validity. Finally, we offer guidelines for practical implementation and conclude with a discussion of the limitations of the twin framework.

Literature review

The use of technology-assisted platforms to aid in written academic work (with associated impacts on assessment validity) predates GenAI (Prentice & Kinden, 2018; Roe & Perkins, 2022), as do other threats to assessment validity and academic integrity, such as contract cheating (Clarke & Lancaster, 2006). However, the advanced capabilities of GenAI to produce extended works have led to a focus on GenAI-assisted plagiarism, or 'Aigiarism’ (Khalaf, 2025). As

a result, it is challenging to identify whether a student's work is their own. This compromises validity as it becomes impossible to identify whether students have met the required standards for a course (Dawson et al., 2024) and thus fails to provide assurance of learning. Although GenAI tools have been publicly available since 2022 (OpenAI, 2022), no clear answer has emerged to fully resolve this so-called “wicked problem”, only better or worse solutions (Corbin et al., 2025a). AI detection was rapidly promoted as a potential remedy; however, studies have shown that these technologies do not work sufficiently to make informed decisions on student usage (Chaka, 2023; Perkins et al., 2024b; Weber-Wulff et al., 2023), and thus detecting GenAI use in assessments is now “all but impossible” (TEQSA, 2025, p. 2). Furthermore, such surveillance-focused responses may impact the relational dimension of assessment. Carless (2009) highlighted this point, suggesting that trust must be developed between students and institutions for effective assessment reform.

In addition to detection, other strategies have also been proposed. These include relying less on surveillance technologies and more on providing student support (Luo, 2024), embedding AI literacy into higher education curricula (Foung et al., 2024), incorporating self-reflection tasks (Combrinck & Loubser, 2025), abandoning certain assessment types (Kofinas et al., 2025), and embedding contextual learning elements into assessments (Gonsalves, 2025).

Essien et al. (2024) contend that offering clear ethical guidelines may prevent GenAI misuse, while Cotton et al. (2024) suggest that a mixture of approaches, including educating students, requiring multiple draft submissions, using detection tools, and closely monitoring student work, are all potentially effective strategies. Although these strategies are valuable to different extents, their effectiveness remains uncertain. Detection-based strategies may be unreliable (Weber-Wulff et al., 2023; Perkins et al., 2024b) and ‘discursive’ approaches that do not tackle assessment redesign rely on compliance (Corbin et al., 2025b). As a result, structural assessment redesign appears to be a more robust and pragmatic response; however, the literature offers few clear methods for this purpose. Frameworks and approaches for fostering assessment audits and redesign have also been developed, including the PANDORA GenAI Susceptibility Rubric (Bannister et al., 2025), AI Assessment Scale (AIAS) (Perkins et al., 2025a; Perkins et al., 2024a), ‘traffic light’ systems to communicate acceptable GenAI usage (University of Leeds, 2025), and ‘lanes’ for secured and unsecured assessments with and without access to GenAI (Bridgeman et al., 2024). It has been argued that approaches that only communicate guidelines without accompanying structural changes (discursive changes) are inadequate for addressing GenAI in assessment (Corbin et al., 2025b). Likewise, misapplications of frameworks such as the AIAS without due consideration of structural change may also cause unintentional harm (Perkins et al., 2025b). Much like the other options discussed above, we do not believe that assessment change can fully resolve the issues brought forward by AI – but we contend that it is a useful method for enhancing assessment validity. At the same time, we argue that existing forms of assessment (such as take-home essays, portfolios, and unsupervised, authentic pieces of work) still have legitimate value and a rightful place in summative assessment protocols. In reference to essays specifically, we concur with Corbin et al.’s (2026, p. 202) point that essays have a unique capacity to develop intellectual abilities that cannot be cultivated as effectively through other means, including self-regulated learning and metacognitive awareness. The educative value of these forms of assessment is at the core of the philosophy behind creating assessment twins.

When we discuss validity as part of an assessment-twin protocol, we frame our understanding through Messick's (1989, 1993) seminal works. Traditional conceptions of assessment validity are classified into three types: content, criterion-related, and construct validity. Messick (1989, 1993) challenged this assertion and proposed a unified model in which construct validity serves as the overarching framework that subsumes all other validity aspects. According to this model, validity can be defined as an evaluative judgment on the extent to which evidence supports the appropriateness, meaningfulness, and usefulness of assessment results. It is important to note that validity judgments are made about specific assessment tasks and their results, not about general approaches to assessment (Messick, 1989). Accordingly, in this paper, we use Messick's framework to analyse the validity implications that arise when specific assignments are designed and implemented as twins; the tables and discussion that follow illustrate the types of validity considerations that specific twinned tasks must address, rather than making a validity claim about twinning as a technique in the abstract. Messick did not explicitly label his framework as consisting of six strands; however, later works (Shaw & Crisp, 2012) have drawn on Messick to frame six sources of validity evidence: content, substantive, structural, generalisability, external, and consequential. Each of these contributes to the overall construct validity, as shown in Table 1.

Table 1. Shaw and Crisp's (2012) Six-Strands of Validity, based on Messick (1989, 1993).

Validity Strand	Definition
Content	Relates to the representativeness and relevance of the content.
Substantive	Addresses the justifications and theoretical basis for the consistency of the assessment, comparability of the underlying cognitive processes to the assessment and performance in practice.
Structural	Addresses the reliability of the procedures for assigning scores and the scoring processes.
Consequential	Regarding the consequences of the assessment for the person who is taking the assessment.
Generalisability	Addresses the question of to what degree score properties or interpretations can be widened and generalised in different contexts.
External	Describes the relationship between assessment scores and scores on other assessments that measure the same thing.

Assessment twins

The premise behind using assessment twins is that when one form of assessment is more vulnerable to GenAI completion, pairing it with a complementary task that is less vulnerable provides greater assessment validity and a clearer representation of assessment performance. Consequently, we define assessment twins as two deliberately designed, interdependent assessment components that (a) address the same intended learning outcomes, (b) require different modes of evidence or production, and (c) are scheduled so that performance on each component can be cross-checked to mitigate a known vulnerability (e.g., GenAI completion, impersonation), thereby enhancing construct validity compared to either component considered alone.

An assessment twin strategy does not require educators to abandon established assessment types such as essays or reports, which we have argued are pedagogically valuable and difficult to replace. In contrast, an assessment twin acknowledges the role of these assessments but seeks to enhance their validity by gathering additional evidence. Therefore, it is important to clarify why both components of the twin are retained rather than simply replacing the GenAI-vulnerable task with a resistant one. The distinction between formative and summative assessments is central here. Formative assessments are primarily designed to foster learning; they provide opportunities for students to engage with material, develop skills, and receive feedback that informs their progress (Bennett, 2011). Summative assessments, by contrast, are primarily designed to evaluate and certify the final achievement of learning objectives (Craddock & Mathias, 2009). Many GenAI-vulnerable tasks – such as take-home essays or research reports – carry substantial formative value: the process of completing them is itself a meaningful learning activity, regardless of whether the resulting artefact can be fully trusted as evidence of achievement. Discarding such tasks entirely would sacrifice this learning opportunity. Assessment twins, therefore, retain the GenAI-vulnerable task for its formative value while pairing it with a twin that provides reliable summative evidence of the same learning outcomes. In this way, both components serve distinct and complementary roles: the original task supports learning, and the twin confirms achievement.

The twin approach builds on existing, long-established assessment practices, such as the oral viva voce, which is commonly associated with thesis defences in postgraduate assessment. We also foresee an assessment twin protocol that is suitable for summative assessment, in which the objective is to judge learning (Bennett, 2011; Crisp, 2012) or certify achievement (Craddock & Mathias, 2009). However, creating an assessment twin is distinct from the simple process of pairing a written essay with a traditional oral viva voce. Notably, a twin task for a GenAI-susceptible assessment could take multiple formats, including a group interview or peer discussion, a timed in-class test, or the production of a physical artefact. The underlying principle is one of complementary modes of assessment, which promotes authenticity and creates a system of checks and balances in which inconsistencies in understanding, proficiency, or competency come to the surface. Another benefit of twinned assessments is their flexibility. Twin elements can be both low-tech (e.g., in-class discussions, oral defences) or high-tech (for example, creating an in-class concept map as a group using AI tools). The twin approach can also be applied to small classes of a few individuals or larger groups.

How do assessment twins enhance validity?

Messick's (1989) six-strand approach provides a lens for analysing how GenAI disrupts traditional assumptions regarding assessment validity (see Table 2). Each of these strands of validity is now exposed to new, uneven pressures in GenAI-enabled educational contexts. Tasks that previously aligned with certain constructs may now be exposed to shortcuts, or score interpretations may no longer be reliable.

The impact of GenAI on these dimensions of validity is not singular; rather, GenAI may threaten different strands of validity simultaneously. In Table 2, we map the ways in which these six strands of validity are disrupted by the attributes of GenAI models and propose ways in which validity can be strengthened through the redevelopment of assessment through a twin process.

Table 2 illustrates how each strand of Shaw and Crisp's (2012) adaptation of Messick's validity framework is susceptible to disruption by GenAI and how a specific twin strategy can be designed to address that disruption for a given assessment task. Reading across each row, the GenAI threat column identifies the nature of the vulnerability, the twin strategy column proposes a concrete pairing that targets that specific vulnerability, and the final column explains the form of validity evidence that the twin generates. For example, content validity is threatened when a student submits GenAI-generated work without genuine engagement with the material; pairing a take-home task with an in-person discussion of key concepts produces direct evidence of understanding that the written task alone cannot reliably provide. Taken together, the rows demonstrate that no single strand of validity is immune to GenAI disruption and that the appropriate twin design will therefore vary depending on which strand is most at risk in a given assessment context. It is the assessor's responsibility to consider these validity dimensions when designing specific twinned tasks, as validity is a property of particular assessment results rather than of the twin approach itself.

In summary, each of these elements of overall construct validity can be enhanced by adopting a twin approach in the assessment strategy.

Practical design for assessment twins

Although we provide conceptual evidence for the twin concept to enhance assessment validity, it is important that this approach is grounded in practical implementation strategies. As a novel framework, there are no existing empirical cases of a twin strategy in action. However, in proposing assessment twins, we foresee that the protocol would be best implemented through an iterative audit and development process. We propose a three-step process here: beginning with the identification of a vulnerability, followed by the consideration of learning outcomes and the development of a complementary assessment, and then the creation of a marking framework, prior to pilot testing the assessment. These steps are outlined below.

Step 1: Identifying assessment vulnerability

Table 2. Mapping GenAI Validity Threats and Assessment Twinning Responses to Strands of Validity.

Validity Type	GenAI Threat	Assessment Twin Strategy	Validity is Enhanced by
Content Validity	Learners submit work to an assessment designed to evaluate knowledge on an issue using GenAI models, thereby bypassing the learning process.	Twin take-home assessments with in-person discussions of key concepts.	Confirmation of understanding in relation to learning outcomes.
Substantive Validity	Learners bypass specific cognitive processes or synthesis techniques through GenAI usage (e.g. Using OpenAI's deep research to conduct a literature review).	Explain or document learners' cognitive processes.	Evidence of engagement with required cognitive processes. For example, a demonstration of a core skill.
Structural Validity	The high assessment scores of GenAI-produced content undermine score validity.	Link scores across assessment elements, for example, cap scores on a written task if oral explanation is poor.	Range of performance verified across subjects to maintain score reliability.
Consequential Validity	The assessment encourages a surface-level approach and induces dependency on GenAI tool usage. This detracts from learner agency and autonomy in the long term, thus is a negative consequence induced by GenAI vulnerability.	Incorporate a twin that has a metacognitive element (i.e., an oral self-assessment) and ask students to reflect on their use of GenAI.	Promotes self-regulation, critical awareness of GenAI impacts, and stimulated learner reflection.
Generalisability	Learners perform well only in environments where they have technology access but fail to replicate performance when they do not have GenAI access.	Improve generalisability of results by assessing the same set of learning outcomes in low- and high-technology contexts, with restrictions on AI use where necessary.	Assessing the ability to apply knowledge in GenAI and non-GenAI-enabled contexts.
External Validity	Assessment scores are unreliable indicators of real-world performance.	Incorporate practice-based or scenario tasks for human-confirmed elements.	Triangulation of applied and non-applied skills.

The first step toward creating assessment twins is to identify whether existing assessments are threatened by GenAI capabilities. This requires assessors to have a threshold level of AI literacy, for example, by understanding the strengths and limitations of current GenAI models and what they can and cannot do. Broadly speaking, if assessment outcomes are threatened by the production of high-quality GenAI output with little to no human input, there is a strong argument that the validity of the assessment is challenged. Tools such as the PANDORA rubric (Bannister et al., 2025) can be valuable in this part of the process. Additional considerations include the context in which the assessment takes place: if students can undertake the assessment remotely without human observation, supervision, or invigilation (i.e., proctored or supervised examination conditions), there is a greater likelihood that the validity of the assessment will be threatened. Additionally, the ease with which GenAI content can be differentiated from human work may be a deciding factor. For example, an art project undertaken using canvas and oil paints would meet the criteria of being achievable or completable remotely without supervision; yet, it would not be meaningfully vulnerable to GenAI completion. If assessments fulfil most or all of these criteria, requiring a twin assessment may help maintain validity.

Step 2: Consider learning outcomes and choose a twin assessment

The next step is to define and explicate the learning outcomes that the assessment is required to measure. This includes any competencies, skills, or knowledge required to pass the assessment. Assessment twins should not be two disconnected tasks, and the value of the approach lies in the fact that both components should map onto the same set of learning outcomes in a complementary way. By focusing on the intended learning outcomes, educators can identify which dimensions of learning are most likely to be compromised by GenAI.

There are multiple ways to design assessment twins. The exact format of a successful twin depends on the learning context, institutional requirements, resource constraints, and the nature of the subject being assessed. Complementary modes to traditional written assessments could include real-time demonstrations; for example, oral explanations, group discussions, or question-and-answer sessions. However, this may not be a feasible option in resource-limited contexts with large student cohorts. In such cases, peer assessment, for example, team-based assessment, could be explored. The twin should be less vulnerable to GenAI completion while retaining the measurement of the intended learning outcomes.

For example, if a learning outcome relates to critically evaluating source material, a written essay may be useful to demonstrate clear, structured arguments, whereas an oral discussion or video recording of a reflection on the work may help verify students' reasoning processes. As a second example, the application of knowledge in practice: in this case, a simulation of an authentic task or an in-class problem-based task could be combined with a secondary written report.

It should be noted that some assessment formats that appear less susceptible to GenAI may still carry vulnerability if they rely on students' unsupervised preparation. For instance, oral presentations and video submissions are only as valid as the work that underpins them; if students can generate scripts or talking points using GenAI and simply read or recite these during a presentation, the assessment does not reliably evidence independent understanding. Similarly, group projects may be vulnerable if individual contributions are produced with AI assistance. To mitigate these risks, twin tasks that involve oral or performance-based elements should incorporate live, interactive components, such as unscripted questions, real-time discussions, or in-situ problem solving, that make it difficult for students to rely on pre-generated AI content.

Step 3: Develop a marking framework

A key principle behind an assessment twin is the interdependence between the two tasks. This does not mean that a grade or weighting is assigned for one component and the other (i.e., a 50% weighting on a pre-prepared presentation and a 50% weighting on an interview). The marking approach for the assessment requires careful consideration. This could include a confirmatory aspect; for example, performance in the second assessment is required to confirm performance in the first assessment (i.e., it is a 'yes/no'). A threshold may also be established to suggest a minimum performance on the twin to validate the GenAI vulnerable assessment.

To illustrate how these principles might work in practice, consider the following example. A student submits a take-home research essay (the GenAI-vulnerable task) and subsequently completes a 15-minute oral interview in which they are asked to explain their argument and respond to questions about their sources (the twin). The essay is marked out of 100 using a standard rubric. The oral interview is then used in one of two ways, depending on the assessor's chosen approach. Under a threshold model, the student must demonstrate a minimum level of understanding in the oral interview – for example, they must be able to accurately explain at least two of the key arguments in their essay and respond coherently to at least one follow-up question – for the essay mark to stand. If they fall below this threshold, the essay is not awarded credit, regardless of its written quality. Under a confirmatory weighting model, the oral interview is scored on a simple scale (e.g., 0, 1, or 2) reflecting the degree to which it confirms the essay. The final mark might then be calculated as: essay mark × (oral confirmation score / 2). In this case, a student who scores 80 on the essay but receives a 1 out of 2 on the oral would receive a final mark of 40, reflecting the partial confirmation of their performance. It should be noted here that using a confirmatory weighting model will result in a score of zero on an oral component, effectively rendering the essay mark irrelevant; therefore, educators must consider institutional and contextual considerations when establishing whether this is appropriate in their assessment design.

We recognise that there are contexts in summative higher education assessments where assessment twins may not be appropriate, as shown in Table 3.

Table 3. When (not) to use twins.

Use a twin when…	Do not twin when…
A task is pedagogically rich but AI susceptible (essay, take-home coding, design brief).	Outcomes differ across the two tasks: a single redesign (for example, an authentic, supervised studio task) would already be valid and manageable.
The same learning outcomes can be evidenced via a second, lower-risk mode (supervised discussion, rapid in-class derivation, oral walkthrough, process log).	Workload or equity concerns (e.g. time-intensive viva voces for large cohorts) cannot be mitigated using scalable alternatives.
Institutional constraints rule out full-scale supervised examinations but allow other authenticity checks.	High-stakes, single-session exams are feasible and already secured.

Strategies for implementing a twin assessment design

Twin assessment design offers an approach to maintaining validity in a world in which GenAI tools are widely available. However, we recognise that implementing assessment restructuring, such as a twin strategy, is easier said than done and is heavily context specific. In this section, we explore some of the implementation challenges associated with a twin assessment practice. One of the fundamental issues that we anticipate in terms of using twin assessments in higher education is the resource-intensiveness of providing additional, in-person confirmatory tasks (such as oral viva voces), which may not be possible in the context of large class sizes.

We also recognise that there are contexts in which an assessment twin approach will not be appropriate. Therefore, we suggest that assessment twinning be chosen only in specific circumstances, as discussed in Table 3. For example, if an existing assessment is pedagogically valuable yet GenAI vulnerable, this is the most important criterion for implementing a twin strategy. Even assessments potentially vulnerable to GenAI may still retain pedagogical value, and instructors may still wish to retain these as part of a formal assessment task, rather than changing them to a formative assessment or learning activity.

Furthermore, assessment twin strategies are suited to institutions that can support the resources required for implementation, and in contexts where student learning is enhanced by multimodal assessment. In contrast, when dealing with resource-limited contexts or extremely large class sizes, we would argue that a twin approach may still be possible but is not optimal. In these cases, a complete redesign of the overall assessment strategy using an established framework (such as the AIAS) is more likely to yield results in enhancing validity.

Twin assessment strategies for different cohort sizes

Accordingly, we recognise that the administration of assessments and assessment redesign must focus on the realities of cohort sizes. We categorise these as small groups (between 5 and 25 students), medium groups (25–75 students), and large groups (75 and above).

Small group assessment twin strategies (5–25 students)

In a smaller group size, there is greater potential for the instructor to interact personally with each student and develop a relationship, over time, and through formative assessment, potentially understanding the learners' position, capabilities, and areas for development. This lends itself to a less resource-intensive assessment twin design.

Furthermore, a smaller group size requires fewer logistical considerations. In this context, a GenAI vulnerable assessment could be paired with an individual oral examination or viva voce, a small group discussion with rotating student facilitators, a peer review session, or an individual consultation. In terms of implementation, twin components may be scheduled during learning hours or classes or in dedicated assessment time. In this context, a twin assessment approach will provide quality validity evidence.

Medium group assessment twin strategies (25–75 students)

As the size of a cohort increases, so too do the resource requirements for designing and delivering twin assessments. A strategy to mitigate this is to incorporate group assessment formats. Examples of assessments that could be twinned with a GenAI-vulnerable assessment element include peer-group presentations, larger simultaneous group discussions in which the instructor briefly spends time with each group, or poster presentation events. If possible, incorporating multiple assessors may make this approach more viable.

Large group assessment twin strategies (75+ students)

A large group that requires multiple forms of assessment poses resource constraints and significant challenges to implementing twinned assessment practice; however, this remains viable with some caveats. Bearing in mind that validity is always a claim rather than an absolute (Dawson, 2020), there are still benefits to a twinned approach in terms of providing validity evidence. Examples of a twinned assessment strategy that would be effective in enhancing validity for such a group could include a random sampling approach, in which a percentage of students are selected for a detailed twin assessment, or peer group discussion sessions with multiple assessors (if practical). Video submissions may be vulnerable to technological manipulation, such as deepfakes (Runyon, 2025); however, they could still provide another significant data point in collaboration with other, more significantly GenAI-vulnerable tasks (such as a take-home, written assessment). In-person examinations remain an important and secure form of assessment and could be part of a twinning approach if the same outcomes as the GenAI-vulnerable assessment are being assessed.

When a random sampling approach is adopted for large cohorts, questions of fairness inevitably arise. To address this, it is essential that all students in the cohort are equally eligible for selection and that the sampling is genuinely random, rather than based on suspicion or prior performance. Students should be informed in advance that a proportion of the cohort will be selected for the twin assessment and that selection will not affect their standing or grades unless the twin reveals a discrepancy with the original task. This transparency helps maintain trust and ensures that the approach is perceived as equitable rather than punitive.

Limitations and future research

Implementation and scalability

The most pressing challenge in implementing assessment twins in higher education is resource intensity: this approach requires faculty time, administrative coordination, and institutional support. For large student cohorts, scalability becomes especially problematic. Although we have proposed solutions, such as random sampling, these could compromise the validity that the twin strategy aims to safeguard.

We also face complex administrative barriers, including scheduling logistics, maintaining consistent scoring across multiple assessors, and reconciling grades when twin components produce conflicting outcomes. Even if we can overcome these hurdles, quality assurance may remain uncertain. We must therefore acknowledge that while assessment twins have value, they may not yet be practical in some educational contexts, especially those already facing resource constraints.

Equity and accessibility

We must also confront serious questions of fairness. As assessment research warns, design choices that overlook inclusivity can unintentionally deepen inequities (Lynam & Cachia, 2018). Assessment twins may disadvantage students with diverse communication styles, social anxiety, or cultural backgrounds that make oral examinations and group discussions especially challenging. While this may pose difficulties for students, it remains an unavoidable and important aspect of higher education. To support learners in developing this skill, educators can offer strategies for developing confidence and competence in public speaking, provide sufficient practice and scaffolding, and guide students on how to deliver a presentation effectively. In addition, Universal Design for Learning (UDL) principles may be useful as part of the redesign process. Using UDL principles may help those engaged in redesign to anticipate diverse learner needs, thereby reducing the risk of needing to retrofit accommodations. For example, a UDL approach can help educators consider emotional variability and learner diversity, which may help deliver more meaningful and accurate assessment results (Rose et al., 2018).

Students with disabilities may face additional barriers if accommodations are not considered across both components. Meanwhile, an increased assessment load risks placing unequal pressure on students who juggle family responsibilities, employment, or limited study time. We also recognise that language barriers may unfairly affect international students or those for whom English is not a first language, particularly in oral formats. Perhaps most concerning, if we frame the approach primarily around catching GenAI misuse, we risk fostering a surveillance mentality that positions students as potential cheaters rather than learners (Giray, 2024b; Dawson, 2020), undermining the trust essential to education. Without deliberate attention to inclusive design, we may unintentionally create more inequitable learning environments rather than fairer ones.

Empirical limitations

We must acknowledge that to date, no empirical data exists on the effectiveness of this approach. While the framework assumes that inconsistencies between twin components provide assurance of learning by identifying discrepancies between performance in each assessment task, we recognise that such inconsistencies could reflect legitimate factors such as anxiety, uneven skill development, or differences in comfort with assessment formats (Struyven et al., 2005), and different assessment formats may capture different learning outcomes rather than equivalent ones (Shaw & Crisp, 2012). Elshall and Badir (2025) have called for hybrid approaches that combine traditional methods with AI-assisted projects; however, such models remain largely unexplored. Despite these significant limitations, we believe that assessment twins have utility. Engaging in a twin process forces us to grapple with urgent questions about validity, fairness, and trust in an era of rapidly evolving GenAI tools. In this sense, assessment twins should not be viewed as a perfect or final solution but as a bold and necessary experiment.

Future research directions

Future research should prioritise the empirical validation of the assessment twin framework through systematic investigation across multiple educational contexts. This work is necessary for refining the framework and establishing its practical value. Key priorities should include controlled pilot studies in small-scale settings to test the effectiveness of providing triangulation through multiple data points, mixed-methods research that captures both quantitative outcomes (such as grade correlations) and qualitative experiences (including student stress levels and faculty workload), and longitudinal studies that track whether twin assessments enhance learning and academic performance. Future research should also engage in comparative research that evaluates assessment twins alongside alternative strategies, such as authentic assessment reforms (Crisp, 2012) or AI-integrated pedagogies (Foung et al., 2024). Importantly, it must be explored whether twin performance links meaningfully to real-world competencies, providing a stronger basis for claims of predictive validity. Among institutions and educators who explore using assessment twins, it is important to establish ways of monitoring impact and effectiveness. This could include examining correlations between grades of twin components, recording any outliers or discrepancies, and ensuring that student feedback is part of the development process.

Conclusion

In this study, we propose the use of assessment twins to address the challenges created by GenAI in higher education. Our framework supports the validity of the assessment while preserving the pedagogical value of established tasks. By pairing two assessments that address the same learning outcomes through different modes of evidence, we provide opportunities for cross-verification and generate stronger claims about the assurance of learning.

We argue that validity is strengthened when multiple tasks converge on the same outcome, and we emphasise that the adaptability of twins across different cohort sizes makes this approach widely relevant. At the same time, we recognise the limitations. Assessment twins demand extra resources, thoughtful workload management, and inclusive design to avoid inequities. Without careful planning, risks such as stress, inefficiency, or superficial adoption may undermine the potential of a twin approach. We therefore call for further research, experimentation, and refinement to test and improve the model. Despite these challenges, we contend that this twin assessment strategy may be one of many methods that can support the enhancement of the validity of assessments and support learning in the new GenAI era. Assessment in higher education must move beyond surveillance-driven practices toward a model in which multiple assessment modes become complementary learning opportunities (Giray et al., 2025). In this sense, twins align with broader movements in higher education that call for more holistic, authentic, and student-centred approaches.

Acknowledgements

We are grateful for the ideas contributed by Leon Furze and Thomas Corbin in the initial development phases of this piece.

References

Bannister, P., Urbieta, A. S., & Alvira, N. B. (2025). Appraising higher education assessment validity: Development of the PANDORA GenAI susceptibility rubric. *Journal of Applied Learning & Teaching, 8*(1), 41–55. https://doi.org/10.37074/jalt.2025.8.1.20

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bridgeman, A., Liu, D., & Weeks, R. (2024, September 12). *Program level assessment design and the two-lane approach*. Teaching@Sydney. https://educational-innovation.sydney.edu.au/teaching@sydney/program-level-assessment-two-lane/

Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education, 34*(1), 79–89. https://doi.org/10.1080/02602930801895786

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching, 6*(2), 1–11. https://doi.org/10.37074/jalt.2023.6.2.12

Clarke, R., & Lancaster, T. (2006, June). *Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites* [Paper presentation]. 2nd International Plagiarism Conference, Newcastle, UK.

Combrinck, C., & Loubser, N. (2025). Student self-reflection as a tool for managing GenAI use in large class assessment. *Discover Education, 4*(1), Article 72. https://doi.org/10.1007/s44217-025-00461-2

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025a). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2025.2553340

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

Corbin, T., Walton, J., Bannister, P., & Deranty, J.-P. (2026). On the essay in a time of GenAI. *Educational Philosophy and Theory, 58*(3), 198–210. https://doi.org/10.1080/00131857.2025.2572802

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Craddock, D., & Mathias, H. (2009). Assessment options in higher education. *Assessment & Evaluation in Higher Education, 34*(2), 127–140. https://doi.org/10.1080/02602930801956026

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education, 37*(1), 33–43. https://doi.org/10.1080/02602938.2010.494234

Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. https://doi.org/10.4324/9780429324178

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

Department for Education. (2025, August 12). *Generative artificial intelligence (AI) in education*. GOV.UK. https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education

Elshall, A. S., & Badir, A. (2025). Balancing AI-assisted learning and traditional assessment: The FACT assessment in environmental data science education. *Frontiers in Education, 10*, Article 1596462. https://doi.org/10.3389/feduc.2025.1596462

Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education, 49*(5), 865–882. https://doi.org/10.1080/03075079.2024.2316881

Foung, D., Lin, L., & Chen, J. (2024). Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence, 6*, Article 100250. https://doi.org/10.1016/j.caeai.2024.100250

Giray, L. (2024a). “Don’t let Grammarly overwrite your style and voice”: Writers’ advice on using Grammarly in writing. *Internet Reference Services Quarterly, 28*(3), 293–303. https://doi.org/10.1080/10875301.2024.2344762

Giray, L. (2024b). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *The Serials Librarian, 85*(5–6), 181–189. https://doi.org/10.1080/0361526X.2024.2433256

Giray, L. (2025). When using AI in scientific research: Start with human, end with human. *TechTrends, 69*(1), 1–8. https://doi.org/10.1007/s11528-025-01132-7

Giray, L., De Silos, P. Y., Adornado, A., Buelo, R. J. V., Galas, E., Reyes-Chua, E., Santiago, C., & Ulanday, M. L. (2024). Use and impact of artificial intelligence in Philippine higher education: Reflections from instructors and administrators. *Internet Reference Services Quarterly, 28*(3), 315–338. https://doi.org/10.1080/10875301.2024.2352746

Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college writing. *Internet Reference Services Quarterly, 29*(1), 83–116. https://doi.org/10.1080/10875301.2024.2437174

Gonsalves, C. (2024). Generative AI’s impact on critical thinking: Revisiting Bloom’s taxonomy. *Journal of Marketing Education*, *48*(1), 4-19. https://doi.org/10.1177/02734753241305980*,*

Gonsalves, C. (2025). Contextual assessment design in the age of generative AI. *Journal of Learning Development in Higher Education, 34*. https://doi.org/10.47408/jldhe.vi34.1307

Khalaf, M. A. (2025). Does attitude towards plagiarism predict aigiarism using ChatGPT? *AI and Ethics, 5*(1), 677–688. https://doi.org/10.1007/s43681-024-00426-5

Kofinas, A. K., Tsay, C. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology, 56*(2), 465–486. https://doi.org/10.1111/bjet.13585

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the “originality” of students’ work. *Assessment & Evaluation in Higher Education, 49*(5), 651–664. https://doi.org/10.1080/02602938.2024.2309963

Lynam, S., & Cachia, M. (2018). Students’ perceptions of the role of assessments at higher education. *Assessment & Evaluation in Higher Education, 43*(2), 223–234. https://doi.org/10.1080/02602938.2017.1329928

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series, 1993*(2), i–18. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2). https://doi.org/10.53761/1.20.02.07

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024a). 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), Article 6. https://doi.org/10.53761/q3azde36

Perkins, M., Roe, J., & Furze, L. (2025a). Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice, 22*(7). https://doi.org/10.53761/rrm4y757

Perkins, M., Roe, J., & Furze, L. (2025b). How (not) to use the AI Assessment Scale. *Journal of Applied Learning & Teaching, 8*(2), 14–23. https://doi.org/10.37074/jalt.2025.8.2.15

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024b). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*(1), Article 53. https://doi.org/10.1186/s41239-024-00487-w

Prentice, F. M., & Kinden, C. E. (2018). Paraphrasing tools, language translation tools and plagiarism: An exploratory study. *International Journal for Educational Integrity, 14*(1), Article 11. https://doi.org/10.1007/s40979-018-0036-7

Roe, J. (2025). *How to use generative AI in educational research*. Cambridge University Press. https://doi.org/10.1017/9781009675338

Roe, J., & Perkins, M. (2022). What are automated paraphrasing tools and how do we address them? A review of a growing threat to academic integrity. *International Journal for Educational Integrity, 18*(1), Article 1. https://doi.org/10.1007/s40979-022-00109-w

Roe, J., & Perkins, M. (2024). Generative AI and agency in education: A critical scoping review and thematic analysis. *arXiv*. https://doi.org/10.48550/arXiv.2411.00631

Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and informative for all: Universal design for learning (UDL) and the future of assessment. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of Accessible Instruction and Testing Practices: Issues, Innovations, and Applications* (pp. 167–180). Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3_11

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Rudolph, J., Tang, F. X., Aspland, T., & Stafford, V. (2025). What does 'good teaching' mean in the AI age? *Journal of Applied Learning & Teaching, 8*(2), 6–13. https://doi.org/10.37074/jalt.2025.8.2.1

Runyon, N. (2025, February 20). *Deepfakes on trial: How judges are navigating AI evidence authentication*. Thomson Reuters Institute. https://www.thomsonreuters.com/en-us/posts/ai-in-courts/deepfakes-evidence-authentication/

Shaw, S., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters, 14*, 2–7. https://doi.org/10.17863/CAM.100449

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 325–341. https://doi.org/10.1080/02602930500099102

Tertiary Education Quality and Standards Agency. (2025, September 24). *Enacting assessment reform in a time of artificial intelligence*. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/enacting-assessment-reform-time-artificial-intelligence

University of Leeds. (2025). *Categories of assessments | Generative AI*. https://generative-ai.leeds.ac.uk/ai-and-assessments/categories-of-assessments/

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*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z