
ASSESSMENT TWINS: A PROTOCOL FOR AI-VULNERABLE SUMMATIVE ASSESSMENT

A PREPRINT

Jasper Roe ^{1*}, Mike Perkins ², Louie Giray ³,

¹Durham University, United Kingdom

²British University Vietnam, Vietnam

³Mapúa University, Philippines

* Corresponding Author: jasper.j.roe@durham.ac.uk

Oct 2025

Abstract

Generative Artificial Intelligence (GenAI) is reshaping higher education and raising pressing concerns about the integrity and validity of higher education assessment. While assessment redesign is increasingly seen as a necessity, there is a relative lack of literature detailing what such redesign may entail. In this paper, we introduce *assessment twins* as an accessible approach for redesigning assessment tasks to enhance validity. We use Messick's unified validity framework to systematically map the ways in which GenAI threaten content, structural, consequential, generalisability, and external validity. Following this, we define 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 and assurance of learning.

We argue that the twin approach helps mitigate validity threats by triangulating evidence across complementary formats, such as pairing essays with oral defences, group discussions, or practical demonstrations. We highlight several advantages: preservation of established assessment formats, reduction of reliance on surveillance technologies, and flexible use across cohort sizes. To guide implementation, we propose a three-step 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.

Keywords: assessment twin, generative artificial intelligence, assessment design, assessment validity, higher education

Introduction

Generative Artificial Intelligence (GenAI) remains a crucial area of research in assessment practice in higher education. Since the public release of advanced GenAI models, concerns regarding academic integrity have risen sharply, as the production of stylistically and grammatically coherent text is now widely available (Giray, 2024a; Perkins, 2023). Given that many forms of higher educational assessment have traditionally relied on formats such as the take-home essay, written research project, or pre-prepared presentation, which can now be completed by GenAI, the potential effects on assurance of learning are troubling.

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). Current discourse demonstrates significant uncertainty as to whether GenAI will ultimately help or hinder assessment practices, with significant concern remaining around student use of GenAI in completing tasks. Critics argue that such reliance may lead to reduced learner agency (Roe & Perkins, 2024), dependency, and erosion of critical thinking skills (Giray, 2025; Gonsalves, 2024). These contradictory perspectives underscore the unsettled nature of current educational debates: GenAI presents both opportunities and risks, inviting innovation and threatening core pedagogical aims.

In this paper, we outline a practice for redesigning assessments in higher education in light of the capabilities of developing GenAI models. Specifically, we propose the concept of *assessment twins* for GenAI-vulnerable tasks. In practice, this 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 (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, p1).

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

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). However, the advanced capabilities of GenAI to produce extended works has led to a focus on GenAI-assisted plagiarism, or ‘Aigiarism’ (Khalaf, 2025). As a result, it is challenging to identify whether a students’ 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. Despite the introduction of GenAI tools into the educational landscape several years ago, no clear answer has emerged to resolve this so-called ‘wicked problem’ (Corbin, Bearman, et al., 2025). AI detection was rapidly

promoted as a potential remedy, but studies have shown that these technologies do not work well enough to make informed decisions on student usage (Chaka, 2023; Perkins, Roe, et al., 2024; Weber-Wulff et al., 2023), and so detecting GenAI use in assessments is now “all but impossible” (TEQSA, 2025, p2). 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.

Outside of detection, other strategies have also been posited. 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 assessment (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.

Frameworks and systems for fostering assessment and audit and redesign have also been developed, including the Assessment-GenAI Susceptibility Rubric PANDORA (Bannister et al., 2025), AI Assessment Scale (Perkins et al., 2025; Perkins, Furze, et al., 2024), ‘traffic light’ systems to communicate acceptable GenAI usage (University of Leeds, 2025), and ‘lanes’ for secured and unsecured assessment with and without access to GenAI (Bridgeman et al., 2024). It has however, been argued that approaches which only communicate guidelines without accompanying structural changes (discursive changes) are inadequate for dealing with GenAI in assessment (Corbin et al., 2025).

While assessment redesign is not a silver bullet for addressing the issue of GenAI in assessment in higher education, we contend that it is a valuable method for enhancing assessment validity. At the same time, we contend that existing forms of assessment (such as take-home essays, portfolios, and unsupervised, authentic pieces of work) still have legitimate value and a place in summative assessment protocols. This is at the core of the philosophy behind creating assessment twins.

We frame our understanding of validity through Messick’s (1989, 1993) work on validity. Traditional conceptions of assessment validity are classified into three types: content, criterion-related, and construct validity. Messick (1989, 1993) challenged this assertion, proposing a unified model in which construct validity is the overriding framework under which all other validity aspects are subsumed. According to this model, validity can be defined as an evaluative judgement on the extent to which evidence supports the appropriateness, meaningfulness, and usefulness of assessment results. Messick did not explicitly label his framework as consisting of six strands, but 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	Relates to the justifications and theoretical basis for the consistency of the assessment, how comparable the underlying cognitive processes are vis-à-vis the assessment and performance in practice.
Structural	Relates to how reliable the procedures for assigning scores and scoring processes are.
Consequential	Regards the consequences of the assessment for the person who is taking the assessment.
Generalisability	Asks to what degree can score properties or interpretations be widened and generalised in different contexts.
External	Relates to the relationship between assessment scores and scores on other assessments which measure the same thing.

Assessment Twins

The premise behind an assessment twin 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 validity compared to either component considered alone.

A twin strategy does not require educators to abandon established assessment types such as essays or reports, which can be pedagogically valuable. In contrast, an assessment twin acknowledges the role of these assessments but seeks to enhance their validity by gathering additional evidence. 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 as suited to summative assessment, in which the objective is to judge learning (Bennett, 2011; Crisp, 2012) or certify achievement (Craddock & Mathias, 2009).

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 promote authenticity, creating a system of checks and balances where inconsistencies in understanding, proficiency, or competency come to the surface.

A further benefit of twinned assessments is their flexibility. Twin elements can be both low-tech (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 which previously aligned with certain constructs may now be exposed to shortcuts, or score interpretation may no longer be reliable.

The impact of GenAI on these dimensions of validity is not singular; rather, GenAI may affect several 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 could be strengthened through the redevelopment of assessment through a twin process.

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, effectively bypassing the learning themselves.	Twin take-home assessments with in-person discussion of key concepts.	Confirmation of understanding as it relates 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).	Have learners explain or document their cognitive processes.	Evidence of engagement with required cognitive process. For example, a demonstration of a core skill.
Structural Validity	GenAI produced content receives high assessment scores, undermining score validity.	Link scores across assessment elements, e.g., cap scores on a written task if oral explanation is poor	Cross-verifying range of performance 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 which has a metacognitive element (i.e. an oral self-assessment) and ask students to incorporate a reflection 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	Improve generalisability of results by assessing the same set of learning outcomes in low and	Assessing the ability to apply knowledge in GenAI and non-

	when they do not have GenAI access	high technology contexts, with restrictions on AI use where necessary.	GenAI enabled contexts
External Validity	Assessment scores are not reliable indicators of real- world performance	Include practice-based or scenario tasks for human-confirmed elements	Triangulation of applied and non- applied skills.

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

Practical Design for Twin Assessments

While 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 development of a complementary assessment, followed by the creation of a marking framework, prior to pilot testing the assessment. The proposed steps are outlined below.

Step 1: Identifying Assessment Vulnerability

The first step towards creating assessment twins is to identify whether the existing assessment(s) are threatened by the capabilities of GenAI. This requires the assessor 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 the assessment outcomes are threatened by the production of high-quality GenAI output with little to no human input, then there is a strong argument that the validity of the assessment is challenged. Tools such as the PANDORA rubric (Bannister et al., 2025) may be of value in this part of the process. Additional considerations are the context in which the assessment takes place: if students are able to undertake the assessment remotely, without human observation, supervision, or invigilation, then 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. An art project undertaken using canvas and oil paints, for example, 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, then requiring a twin assessment may help to maintain validity.

Step 2: Consider Learning Outcomes and Choose a Twin Assessment

The following step is to define and explicate the learning outcomes that the assessment is required to measure. This includes any competencies, skills, or knowledge which are 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 that a twin could be designed. The exact format of an assessment twin

depends on the learning context, institutional requirements and resource constraints, and the nature of the subject being assessed. Complementary modes to traditional written assessments could include a real-time demonstration, for example an oral explanation, group discussion, or question and answer session. However, in resource-limited contexts with large student cohorts, this may not be a feasible option. In these cases, forms of peer-assessment could be explored, for example team-based assessment. The twin should be less vulnerable to GenAI completion, while retaining measurement of the intended learning outcomes.

By way of example, if a learning outcome relates to being able to critically evaluate source material, then a written essay may be useful to demonstrate clearly structured arguments, while an oral discussion or video recording of a reflection on the work may help verify the students' reasoning process. A second example could be the application of knowledge in practice: in this case, a simulation of an authentic task or in-class problem based task could be combined with a secondary written report.

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 to the assessment needs careful consideration. This could include a confirmatory aspect, for example the performance in the second assessment is required to confirm the 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.

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). • The same LOs can be evidenced via a second, lower-risk mode (supervised discussion, rapid in-class derivation, oral walk-through, process log). • Institutional constraints rule out full invigilation but allow other authenticity checks.	• Outcomes differ across the two tasks: A single redesign (e.g., authentic, supervised studio task) would already be valid and manageable. • Workload or equity concerns (e.g., time-intensive viva voces for large cohorts) cannot be mitigated with scalable alternatives. • High-stakes, single-sitting exams are feasible and already secure

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 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.

Further to this, assessment twin strategies are suited to institutions which can support the resources required for implementation, and in contexts where student learning is enhanced from multimodal assessment. In contrast, when dealing with resource-limited contexts or extremely large class sizes, then we would argue that a twin approach may still be a possible, but non-optimal solution. 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

In line with the above, we recognise that the administration of assessments and assessment redesign must be focused 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 relatively 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 an assessment that could be twinned with a GenAI-vulnerable assessment element include peer-group presentation, larger simultaneous group discussions in which the instructor briefly spends time with each group, or a poster presentation event. If possible, incorporating multiple assessors may make this approach more viable.

Large Group Assessment Twin Strategies (75+ students)

A large group which requires multiple forms of assessment poses resource constraints and significant challenges to implementing a twinned assessment practice, but this is still 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) yet still could provide another significant data point in collaboration with other, more significantly GenAI-vulnerable tasks (such as a take-home, written assessment). In-person examinations of course 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.

Limitations and Future Research

Implementation and Scalability

The most pressing challenge of 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. While we suggest solutions, such as random sampling, these risk undermining the very validity the twin strategy is designed to protect.

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. Students with disabilities may face additional barriers if accommodations are not considered across both components.

Meanwhile, increased assessment load risks placing unequal pressure on students who juggle family responsibilities, employment, or limited study time. We also recognize 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 recognize that such inconsistencies could just as easily 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, but we note that such models remain largely unexplored. Still, despite these significant limitations, we believe 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 seen as a perfect or final solution, but as a bold and necessary experiment.

Future Research Directions

Future research needs to prioritize the empirical validation of the assessment twins framework through systematic investigation across multiple educational contexts, building on established approaches in assessment research (Boud et al., 2018). This is not simply about testing whether the framework works. Our key priorities should include controlled pilot studies in small-scale settings to test effectiveness in providing triangulation through multiple data points, and 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 actually enhance learning and academic performance. We 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). Just as importantly, we must explore whether twin performance links meaningfully to real-world competencies, giving us a stronger basis for claims of predictive validity. By pursuing these lines of inquiry, we not only fill the current evidence gap but also build a more credible and resilient framework.

Conclusion

In this paper, we have proposed *assessment twins* as a response to the challenges created by Generative AI in higher education. Our framework aims to help ensure validity of 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 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 in enhancing the validity of assessments and supporting learning in the new GenAI era.

Assessment in higher education must move beyond surveillance-driven practices toward a model where multiple assessment modes become complementary learning opportunities (Giray et al., 2025). In this sense, the assessment twins concept aligns with calls for a post-assessment future attentive to power and inequality, where assessment redesigns respond to systemic shifts rather than merely policing tools (Perkins & Roe, 2025).

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., Urbieto, A. S., & Alvira, N. B. (2025). Appraising higher education assessment validity: Development of the PANDORA GenAI susceptibility rubric. *Journal of Applied Learning and Teaching*, 8(1), Article 1. <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>
- Boud, D., Dawson, P., Bearman, M., Bennett, S., Joughin, G., & Molloy, E. (2018). Reframing assessment research: Through a practice perspective. *Studies in Higher Education*, 43(7), 1107–1118. <https://doi.org/10.1080/03075079.2016.1202913>
- Bridgeman, A., Liu, D., & Weeks, R. (2024, September 12). Program level assessment design and the two-lane approach – Teaching@Sydney. *Educational Innovation*. <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, 1–11. <https://doi.org/10.37074/jalt.2023.6.2.12>
- 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. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 0(0), 1–17. <https://doi.org/10.1080/02602938.2025.2553340>
- 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(1), 1–11. <https://doi.org/10.1080/02602938.2025.2503964>
- 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>
- 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., Sevnanayan, 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*, 46(3), 1–15. <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>
- 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), 1–24. <https://doi.org/10.3316/informit.T2024111300009591751711095>
- 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), Article 6. <https://doi.org/10.53761/q3azde36>
- Perkins, M., Jasper, R., & Furze, L. (2025). Reimagining the artificial intelligence assessment scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice*. <https://doi.org/10.53761/rrm4y757>
- Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughan, J., & Khuat, H. Q. (2024). 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., & 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>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), Article 1. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Runyon, N. (2025). 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>
- TEQSA. (2025). *Enacting assessment reform in an age 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>