Talk is Cheap: Why structural assessment changes are needed for a time of GenAI

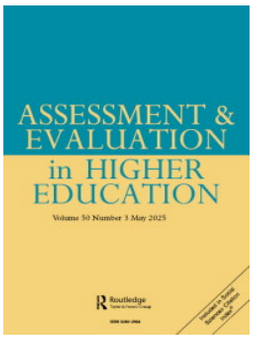
Article *in* Assessment & Evaluation in Higher Education · May 2025

DOI: 10.1080/02602938.2025.2503964

3 authors, including:

Thomas Corbin
Deakin University
19 PUBLICATIONS **181** CITATIONS
SEE PROFILE

Phillip Dawson
Deakin University
120 PUBLICATIONS **8,296** CITATIONS
SEE PROFILE



Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Talk is cheap: why structural assessment changes are needed for a time of GenAI

Thomas Corbin, Phillip Dawson & Danny Liu

To cite this article: Thomas Corbin, Phillip Dawson & Danny Liu (15 May 2025): Talk is cheap: why structural assessment changes are needed for a time of GenAI, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2025.2503964

To link to this article: https://doi.org/10.1080/02602938.2025.2503964

Published online: 15 May 2025.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2025.2503964

Routledge
Taylor & Francis Group

Talk is cheap: why structural assessment changes are needed for a time of GenAI

Thomas Corbin[a], Phillip Dawson[a] and Danny Liu[b]

[a]Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Melbourne, Australia; [b]DVC Education Portfolio, The University of Sydney, Sydney, Australia

ABSTRACT

Generative AI (GenAI) challenges assessment validity by enabling students to complete tasks without demonstrating genuine capability. In response to this challenge, institutions have developed and implemented various approaches that aim to communicate permissible AI use to students. Familiar examples include the 'traffic light' approach now commonly found within institutional policy. While well-intentioned, these approaches share a common limitation: They focus primarily on communicating rules rather than redesigning assessment mechanics. To clarify why such approaches fail, and to guide more effective responses, this paper introduces a novel conceptual distinction between discursive changes to assessment (modifications relying solely on instructions students remain free to ignore) and structural changes (modifications that reshape the underlying mechanics of assessment tasks themselves). Through a critical analysis of prominent frameworks, we demonstrate that current approaches predominantly rely on discursive changes that create what we term an 'enforcement illusion'. We find that educational frameworks frequently borrow the language of socially familiar structural systems (like vehicular traffic lights) while lacking their actual enforcement capabilities, creating an illusion of assessment security. In place of this, we argue for a shift towards structural assessment redesign that builds validity into assessment architecture rather than attempting to impose it through unenforceable rules.

KEYWORDS

Assessment design; academic integrity; artificial intelligence; assessment validity

Generative artificial intelligence (GenAI) increasingly allows students to complete assessment tasks without possessing the relevant knowledge or skills (Nikolic et al. 2024). This presents a fundamental challenge to assessment validity. If assessors cannot determine whether students genuinely possess these capabilities, assessments cease to serve their primary educational purpose (Dawson et al. 2024). In response to this challenge, many academics, scholars, and policy administrators have sought ways to clearly identify and specify to students what uses of GenAI are acceptable within any given assessment context. For example, allowing students to employ GenAI for initial brainstorming but prohibiting its use in actual writing, permitting GenAI to edit language but not to generate substantive content, or explicitly forbidding GenAI assistance in online, unsupervised quizzes. These solutions ostensibly aim to preserve validity by communicating to

CONTACT Thomas Corbin t.corbin@deakin.edu.au Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Melbourne, Australia.

students how and when GenAI can and cannot be used. However, as we will attempt to show in this paper, such approaches rely heavily on student compliance with instructions, a reliance which opens both assessments to vulnerabilities as well as educational institutions to reputational risk.

To address this, in this paper, we introduce and develop a novel conceptual distinction designed to identify the limitations of current responses to GenAI and to provide a sturdy path forward in developing more effective assessment practices. We propose that many of the current assessment frameworks depend primarily on what we term *discursive changes* to assessment, modifications that rely solely on instructing students about permissible AI use, which students remain essentially free to follow or ignore. We contrast these with a different category, which we call *structural changes* to assessment. These are modifications that reshape the underlying mechanics of the assessment tasks themselves, thereby directly influencing or constraining how students can interact with GenAI. Through this distinction, the paper aims to provide a clearer conceptual toolkit for evaluating and redesigning assessments in ways that maintain validity despite the growing capabilities of GenAI.

To achieve the aims of this paper, we first survey several recently emerged and influential frameworks designed to guide assessment redesign in response to GenAI. We discuss prominent approaches such as traffic light systems, AI assessment scales, use-level statements, and multi-tiered classifications, highlighting their intent to clarify permissible AI use to students. After reviewing these examples, we introduce an original conceptual distinction between what we term discursive and structural changes to assessment design. By applying this distinction, we show that existing frameworks predominantly rely on merely discursive methods which introduces significant vulnerabilities related to compliance and enforceability, ultimately undermining assessment validity and institutional reputation. Although these systems may have value in other areas, for example by assisting teachers to conceptualise the different ways AI may be used in a task, from a validity standpoint any change which is merely discursive and not structural is likely to cause more harm than good.

Before introducing existing assessment frameworks, it is important to clarify our scope. As we know, assessment serves multiple purposes (Boud 2000). Assessment *for* learning is important (Carless 2017). However, in this paper we are concerned with moments of assessment *of* learning where there is a desire to in some way control student use of GenAI to enable judgements to be made about student capability. These moments are essential for assessment in higher education to deliver on its promise of ensuring that graduates can do what their university says they can do. With this focus established, we now turn to examine three of the most dominant kinds of AI focused assessment frameworks.

Existing approaches

Assessment design is a complex problem, and educators say they benefit from structures, frameworks and scaffolds to support them (Bearman et al. 2017). As the emergence of GenAI caught many institutions and educators by surprise, there was a rush to develop frameworks that would be useful in this new context. The frameworks that emerged are diverse, spanning straightforward guidelines to complex, scaffolded structures. While they differ in their specific approaches, these frameworks share a common goal which is to specify and clarify appropriate uses of AI in different assessment contexts. In this paper we are concerned with their ability to support assessment validity in assessment of learning. As such, in what follows, we examine the most prominent framework types that have emerged: Traffic light based systems, assessment scales and declarative requirements.

Traffic light systems

Perhaps the most common policy response of higher education institutions to GenAI has been the introduction of 'traffic light' systems to categorise assessments according to permitted levels

of GenAI use. The University of Leeds, as one example among many, employs this three-tiered classification where red translates to 'no AI use permitted', amber to 'limited, assistive AI use allowed', and green allows for 'fully integrated AI use encouraged' (University of Leeds n.d.). This structured traffic light approach is now common internationally.

Macquarie University's implementation provides a detailed illustration of how institutions operationalise traffic light frameworks in practice, offering specific language and structured guidance to help educators communicate permissible GenAI use to students. These instructions aim to 'help you consider levels of acceptable use of generative artificial intelligence tools in assessment tasks' (Macquarie University n.d.). When a teacher chooses the appropriate option from the three familiar-coloured choices, they are then offered prompts to communicate this decision and the correspondingly relevant AI permissions to their students. For example, if they choose Red (meaning 'AI is not permitted'), they are offered the following to communicate: 'For this assessment, students are not permitted to use Generative AI for any purpose, including summarising texts on the subject'. This directive is accompanied by a corresponding Academic Integrity warning for this level; 'Engaging with Generative AI for this assessment constitutes and will be treated as a breach of academic integrity'.

By explicitly labelling permitted and restricted AI engagement levels, universities seek to uphold academic integrity, remove ambiguity around acceptable AI use, and promote transparent dialogue between educators and students (University of Leeds n.d.; University College Dublin, College of Arts and Humanities n.d.). For example, Lancaster University states that their traffic light model, which they refer to as 'RAG' (Red, Amber, Green), has been designed as 'a tool to ensure that for each piece of assessment, staff and students have a shared understanding of whether Gen AI can be used and if permitted how, how much and where in the assessment process' (Lancaster University n.d.) Some institutions, such as the University of Bath, have adopted similar three-tier systems (Type A, B, C, where 'In Type A assessments, the use of any form of GenAI is not permitted'.), which rely on the same traffic light logic (University of Bath 2024; see also University of Reading 2024). However, in all cases, regardless of their naming system, these three-way categorisations are designed to act in the same way, as a framework for clearly communicating to students what is and is not permissible regarding GenAI use.

The AI assessment scale (AIAS)

Developed by Perkins et al. (2024a), the original AI Assessment Scale (AIAS) emerged partly in response to critiques of the Traffic Light model's limitations. While the Traffic Light system is intuitive, Perkins et al. argued that it is too rigid and fails to address the developmental nuances of AI use across a student's academic journey. To address these issues, the original AIAS provided a five-level progression that enables educators to align AI permissions with students' evolving skills and needs. Beginning with No AI Use (Level 1), the scale moves through levels allowing brainstorming, editing, and specific AI-assisted task completion, reaching Full AI Integration (Level 5) in advanced assignments.

To use the AIAS, an educator finds the appropriate level corresponding to the way they want students to use AI in completing an assessment task, and then communicates this to their students. For example, if an educator is aiming to incorporate AI into their assessment, they may want students to use AI to complete certain parts. They would then select Level 4, where 'students are requested or expected to use GenAI to complete specific portions of their tasks, but the emphasis remains on human evaluation and interpretation of the AI-generated content'. Similarly to the traffic light example above, in choosing Level 4 on the scale, the educator will then be prompted to communicate specific instruction to the students. In this case, they will be prompted to tell students; 'You will use AI to complete specified tasks in your assessment. Any AI created content must be cited'. In this way the scale operates by offering different forms of AI use to the educator, allowing them to choose which aligns with the aims of their assessment,

and then offering clear statements they can use to communicate this to students. By differentiating and communicating AI permissions of this kind, Perkins et al. suggest that the AIAS creates an adaptable pathway that reflects students' growing competencies, making it a more nuanced alternative to the Traffic Light model.

A second version of the AIAS was developed in late 2024. The previous five levels were compressed to four, with a new fifth level named 'AI exploration' where students are encouraged to use AI creatively, potentially in co-design situations with their educator (Perkins, Roe, and Furze 2024). In updating the AIAS, the authors write, 'permitting *any* use of AI effectively permits *all* use of AI, and since it is undetectable and sophisticated across domains, the distinction between previous levels 2, 3, 4, and even 5 is somewhat arbitrary'. This is perhaps due to the use of version 2 continuing to depend on student adherence, including statements at different levels such as, students 'may use AI for planning, idea development, and research…', or they 'may use AI to assist with specific tasks such as drafting text…', and 'you must critically evaluate and modify any AI-generated content…'. These directives, as the authors point out, are 'somewhat arbitrary' as there is little reason to think that students will adhere to these instructions and use AI in ways that correspond with the level selected by the educator. Consequently, in revising the AIAS, Perkins, Roe, and Furze (2024) acknowledge that '[r]ather than attempting to control AI use, the AIAS supports transparent conversations between educators and students about appropriate AI use while simultaneously encouraging assessment redesign' (p. 7). In this regard, in an attempt to gain a vision into what students are actually doing when they complete an assessment task, many higher education institutions rely on students declaring their use themselves.

Declarative approaches

Many universities have implemented policies requiring students to declare or disclose their use of AI in assessments. For instance, the University of Melbourne mandates that students acknowledge any AI tools used in their work, specifying the tool's name, the prompts provided, and how the output was utilised (University of Melbourne n.d.). Similarly, Monash University requires students to document their AI use explicitly, often in a dedicated section of their submission, encouraging transparency in how AI tools were employed (Monash University n.d.). Other institutions, such as King's College London, have introduced a mandatory AI use declaration on assessment cover sheets, where students indicate whether they used generative AI and briefly describe its application (King's College London n.d.). However, compliance has been an issue. One study at King's found that up to 74% of students did not complete the AI declaration appropriately, with many fearing that admitting AI use might be perceived negatively (Gonsalves 2024). Many universities, including Cambridge and Princeton take a stricter approach, classifying undeclared AI use as a form of plagiarism and integrating AI disclosures directly into their academic integrity policies, requiring students to explicitly state if AI was used, even for minor tasks like brainstorming (University of Cambridge n.d.; Princeton University n.d.). Some institutions, such as the University of Sydney, allow AI use by default but require students to acknowledge it, treating undisclosed AI assistance as potential misconduct (University of Sydney 2024). Others, such as the University of California, San Francisco, outline specific guidelines for permissible AI use, categorising different levels of acceptable AI involvement and emphasising the need for documentation (University of California, San Francisco n.d.).

Each of these frameworks – traffic light systems, the AI Assessment Scale, and declarative approaches – represent the higher education sector's responses to the challenges posed by GenAI in assessment. While differing in structure and implementation, they all attempt to establish boundaries for appropriate AI use while hoping to maintain assessment validity. In the following section, we present a conceptual framework that allows us to identify the limitations of these approaches.

An alternative approach: the structural/discursive distinction

As we have outlined above, there are many assessment frameworks designed specifically with AI use in mind. Most of these disagree on specifics yet agree that there is a sliding scale of 'appropriateness' which can be identified, articulated, and communicated to students as a means to address the challenges of GenAI. There is, however, an implicit assumption that these levels of appropriateness will be adhered to by students. In fact, within the assessment *of* learning scope of this paper, the success of these approaches is entirely dependent on this adherence. Communication and 'transparent conversations' (albeit important in many ways), accompanied by indications of how AI 'may' or 'should' or 'must' be used, potentially provides both a false sense of security and present a risk to assessment validity. We therefore suggest that any assumption regarding student compliance is problematic. To demonstrate why, it is important to take a step back and understand the options when it comes to assessment design in the age of GenAI.

We suggest that at root there are two possible options for altering assessments with the aim of ensuring validity for a time of GenAI. For any existing assessment, changes can be made either 1) to the way the assessment is communicated to students, or 2) changes can be made to the nature, format, or mechanics of the assessment itself. Naturally, there are many changes possible *within* these two camps, but we suggest that all possible changes will fit within them. We characterise these two options as being 'discursive' in the former case or 'structural' in the latter. In what follows, we will address each in turn.

Discursive changes

We define discursive changes to assessment as:

> *Modifications that rely solely on the communication of instructions, rules, or guidelines to students, such that their success depends entirely on student awareness, understanding, and voluntary compliance with these communications. These changes leave the underlying structure and mechanics of the assessment task unchanged, focusing instead on specifying how students should approach or complete the task.*

A paradigm example of a discursive change to assessment might be as simple as adding 'GenAI use is not permitted in this assessment' to existing instructions. Of course, discursive changes can be far more sophisticated, incorporating detailed rubrics explaining permissible AI use, complex frameworks for different assessment components, or elaborate systems of self-reporting and documentation. However, the sophistication of these changes does not alter their fundamental nature - they remain modifications that work purely through communication and rely on student compliance.

Discursive changes operate through what might be called linguistic commands. They attempt to elicit compliance through language alone, without corresponding mechanisms to enforce those boundaries. In educational contexts, this governance has traditionally been supported by shared values around academic integrity and the implicit threat of detection through plagiarism checkers or stylistic inconsistencies. GenAI fundamentally disrupts this arrangement by making detection unreliable while maintaining the appearance of original work.

The only clear scenario in which purely discursive changes would suffice to adapt existing assessments is also, unfortunately, impractical. This would involve adapting previously suitable assessments – now vulnerable to inappropriate completion by AI – by explicitly prohibiting students from using AI tools. However, this approach would only be viable if reliable AI detection technology existed, which it currently does not (Perkins et al. 2024b; Ardito 2024). False positives and false negatives (Dalalah and Dalalah 2023) continue to be a barrier which are unlikely to be overcome, and alternative approaches to AI detection, such as watermarking, have also been shown to be inadequate (Jiang, Zhang, and Gong 2023). Relying on these unreliable tools harms students through false accusations, while also damaging institutions and teachers by promising

them a level of security these systems simply cannot deliver. Without reliable detection mechanisms, prohibitions against AI use remain merely discursive. This technological limitation exposes a more fundamental issue with discursive approaches. That is, they rely entirely on student compliance with rules that cannot be enforced. Therefore, since perfect detection remains technologically unfeasible, we must examine why discursive approaches are fundamentally vulnerable in practice.

When we examine the assumptions underlying discursive approaches, we find they depend on three increasingly problematic premises. First, discursive approaches assume that students will clearly and consistently understand exactly what is permitted. Yet instructions specifying permissible AI use are inherently ambiguous. Distinctions such as 'AI use for editing' versus 'AI use for drafting' can easily blur into each other, especially when students use multiple rounds of interaction with AI tools. Such ambiguity leaves students uncertain about what constitutes compliance, undermining the clarity which discursive rules attempt to imply. Second, discursive approaches assume that students will voluntarily comply with guidelines, even when non-compliance offers clear practical advantages. This premise not only ignores incentives towards non-compliance but also overlooks genuine disagreements about what constitutes meaningful or legitimate assistance. Recent evidence indicates students' views on appropriate AI uses often diverge significantly from educators' expectations (Corbin et al. 2025). Third, and perhaps most crucially, discursive approaches assume educators have meaningful and reliable mechanisms to verify student compliance. GenAI, however, directly undermines traditional verification methods such as plagiarism detection and stylistic analysis by closely mimicking human writing patterns. As explored above, current detection tools are fraught with false positives and negatives, creating uncertainty and mistrust rather than clarity and accountability. In practice, educators often have no feasible way to confirm whether a student complied with instructions regarding AI use, making compliance essentially unenforceable.

With all this in mind, it seems that although there are many issues with discursive approaches, the main weakness of discursive changes lies not in their clarity or comprehensiveness, but in their enforceability. When an assessment's validity relies purely on students following specific instructions about AI use, we must ask what meaningful mechanisms exist to ensure compliance. The answer is none. This creates what we term the 'discursive paradox': The more detailed and specific our instructions become about 'acceptable' AI use, the more we highlight the gap between what we can specify and what we can verify. Students are increasingly aware of this gap, leaving them to make their own decisions about what counts as appropriate (Corbin et al. 2025). This has consequences which include adding significant burdens to study, creating uneven playing fields between students, and – most importantly for us in this paper – undermining assessment validity.

This paradox ultimately reveals why discursive approaches, while well-intentioned, cannot adequately address the fundamental challenge posed by GenAI to assessment validity. By attempting to control through communication what can only be ensured through structure, these approaches create assessment environments where compliance becomes optional and potentially disadvantageous to students who follow the rules. The alternative, as we will explore next, lies in structural changes that build validity into the assessment design itself rather than trying to impose it through unenforceable rules.

Structural changes

We define structural changes to assessment as:

> *Modifications that directly alter the nature, format, or mechanics of how a task must be completed, such that the success of these changes is not reliant on the student's understanding, interpretation, or compliance with*

instructions. Instead, these changes reshape the underlying framework of the task, constraining or opening the student's approach in ways that are built into the assessment itself.

The power of structural changes lies in their independence from voluntary student compliance. Rather than asking students to follow communicated rules about AI use, structural changes create assessment environments where the desired behaviour emerges naturally from the assessment design. This approach acknowledges that in an age of ubiquitous AI access, assessment validity must be built into assessment structures rather than imposed through guidelines.

To understand how structural changes work in practice, consider the distinction between synchronous and asynchronous assessment. A traditional take-home essay (asynchronous) provides students with ample opportunity to use AI without detection, regardless of what instructions are provided. In contrast, a supervised in-class writing exercise (synchronous) inherently limits AI assistance by its very structure. The student must demonstrate their capabilities in real time, under conditions where external AI assistance is physically restricted. This structural change maintains assessment validity not by asking students to voluntarily limit their AI use, but by creating conditions where inappropriate AI use becomes difficult or impossible. This doesn't mean that all assessment should become synchronous and supervised; certainly, asynchronous assessment has valuable benefits for developing certain skills. The key is aligning the assessment structure with what we genuinely want to measure. If we want to develop a student's ability to think deeply and develop complex arguments over time, an asynchronous format may be appropriate, but we would need to build in structural assessment elements that capture the development process rather than just the final product. The following table provides three additional examples of how existing assessment tasks might be discursively and structurally changed to address AI.

The examples in Table 1 illustrate that structural changes can take many forms, but they share a common principle: they modify the assessment task itself rather than relying on the instructions surrounding it. This might mean shifting from product-focused to process-focused assessment, incorporating real-time demonstration of skills, or creating assignments that leverage the quality of AI interaction rather than attempting to prohibit or limit it. The goal is not to control AI use through rules but to design assessments that are not invalidated by the AI that is likely to be available to students when completing them. We make no claim that the examples in Table 1 represent perfect assessment security; we include them to show how commonly used approaches might fit within our structural/discursive binary.

It is perhaps important at this point, in order to clarify our distinction, to reiterate that our characterisation of 'discursive' hinges on a specific criterion: they rely *solely* on the communication of instructions, rules, or guidelines to students, with no accompanying changes to the underlying mechanics of assessment tasks. This distinction is not merely semantic but fundamental to our argument. Of course, all assessment changes, including structural ones, require communication with students. The difference lies in what is being communicated. Discursive changes communicate rules about how students should approach a structurally unchanged assessment task, while structural changes communicate the nature of a fundamentally altered task. In this sense, there is no meaningful 'hybrid' category within our framework; an assessment change that alters the task mechanics is, by our definition, structural, even though it necessarily involves communication. The critical question is not whether instructions are provided, but whether the

Table 1. Examples of discursive and structural changes to assessment.

	Discursive change	Structural change
Traditional take-home essay	Telling students to use AI for editing but not for generating text	Supervising the generation of parts of the essay
Online multiple choice quiz	Warning screen on first page of quiz telling students to not use AI	Discussing random questions with each student in interactive oral assessment
Lab report	Raising the importance of not fabricating data with AI	Checkpoint in live assessment requiring tutor signoff on lab work

assessment's validity depends entirely on students following those instructions or, moreover, whether the assessment's validity is established by structural task modifications.

After establishing the distinction between structural and discursive changes, we can now turn to critically examine why existing and popular frameworks, such as traffic light systems, ultimately fail to address the fundamental challenge posed by GenAI. While these frameworks appear to offer clarity and structure for managing AI in assessment, they share a common limitation that undermines their effectiveness. This limitation stems from what we call the 'enforcement illusion'.

The enforcement illusion: why discursive approaches fail

The proliferation of frameworks intended to classify and regulate AI use in assessment reveals a common assumption: that the challenge of GenAI can be addressed through clear communication of rules and expectations. This assumption merits critical examination. While frameworks such as traffic light systems and assessment scales appear to offer clarity and structure, they share a fundamental limitation that undermines their effectiveness. These approaches, regardless of their complexity or intuitive appeal, operate entirely through communication rather than enforcement.

Traffic light, scale, dial, or similar models of assessment change are, of course, metaphors. This is no accident. When faced with novel challenges that demand immediate action, metaphors serve as powerful and comforting cognitive tools, helping us navigate unfamiliar territory by mapping it onto known domains. They tell us: 'though this all seems alien and you feel lost, think of it this way and you'll be okay.' Unlike arguments by analogy, metaphors cannot be true or false – they can only be apt or inapt, and as such their value lies not in their truth conduction but in their utility for action. This makes it crucial to examine not whether such metaphors are 'correct', but whether they illuminate or obscure the actual dynamics at play.

The traffic light metaphor is particularly worth examining because its seeming aptness conceals a crucial misunderstanding. That is, real traffic lights work precisely because they are *structural* interventions embedded within robust systems of justification and enforcement. They operate within a physical and institutional infrastructure: cameras automatically detect violations, police regularly patrol intersections, and significant penalties await those who disregard them. Importantly, the dangers of not following the lights are also apparent to drivers. The crucial point is that when authorities identify dangerous intersections, they don't simply post guidelines about when drivers should stop – they install physical infrastructure that forces compliance and enables enforcement. Clearly communicating with drivers is important, but traffic lights don't work just because they are visible.

This reveals the fundamental flaw in using traffic lights as a metaphor for AI policies in assessment. Real traffic lights work precisely because they are structural changes, not discursive ones. In contrast, the above 'traffic light' or derivative frameworks for AI use remain purely discursive – they only communicate rules, not enforce them. The metaphors seeming aptness actually highlights its deep inadequacy. We are borrowing the language of structural change to describe what are merely discursive. This creates a dangerous illusion of control and safety, as educators might assume these frameworks carry the same force as actual traffic lights, when in fact they lack any meaningful enforcement mechanism. The metaphor is not apt. Indeed, it breaks down exactly where it seems most fitting. This helps reveal why we need appropriately structural approaches to assessment design in the age of AI. To unpack this further, we will now turn to some alternative structural strategies.

Structural views into assessment

The distinction between discursive and structural changes suggests that our focus should shift. Rather than asking what rules we should set about AI use, we should ask how we can design

assessments that maintain their validity in an AI-enabled world. This shift in thinking opens new possibilities for assessment design while acknowledging the practical limitations and far-reaching implications of rule-based approaches.

One immediate benefit of adopting this structural perspective is that it provides a clearer lens for evaluating emerging institutional frameworks, such as the University of Sydney's 'two-lane approach' (Liu and Bridgeman 2023). This framework distinguishes between 'Secure' (Lane 1) assessments which are conducted in-person with controlled conditions, and 'Open' (Lane 2) assessments where AI use is uncontrolled (Tertiary Education Quality and Standards Agency 2024, p. 51). The structural/discursive distinction we propose offers a potentially useful lens for understanding and extending the efficacy of such approaches. While Lane 1 assessments incorporate structural elements by creating environments where inappropriate AI use is physically restricted, the effectiveness of Lane 2 assessments depends on how they are designed structurally, as simply designating an assessment as 'Open' without reconsidering its structural mechanics perpetuates the enforcement illusion we have identified. The most effective implementations of dual-track approaches such as these will therefore be those that recognise the need for structural reconsideration of assessment design in both lanes, albeit in different ways. But what might this structural reconsideration look like in practice? While precise structural changes will vary significantly by discipline, learning outcomes, and specific tasks, two broad strategies have emerged as particularly valuable across contexts.

First, structural changes frequently involve reorienting assessment from output to process. Rather than evaluating only the final product, which could potentially be AI-generated, assessment may be designed to capture the student's development and attainment of understanding and skill over time. This might mean building in authenticated checkpoints where students must demonstrate their evolving thinking. For instance, rather than simply submitting a final essay, students might need to participate in live discussions about their developing ideas or demonstrate how their thinking evolved through structured peer feedback sessions. It is important to note however that these sorts of changes likely result in the assessment of different learning outcomes.

Second, structural changes often involve viewing assessment validity at the unit or module level rather than the task level. Instead of trying to ensure each individual assignment is AI-proof (an increasingly futile endeavour), educators can design interconnected assessments where later tasks explicitly build on a student's earlier work. Viewing validity as 'an argument-based evidentiary-chain' (St-Onge et al. 2017, p. 857), the validity of assessment comes not from any single component but from the coherent demonstration of learning across multiple appropriately designed touchpoints.

Unfortunately, as much as researchers or institutions would like to present a clear prescriptive account of the precise structural changes which will allow assessments to remain (or become) valid, that is not a realistic hope. Validity means that an assessment represents the capacity of the student to do or know something. But what that 'something' is changes not only from subject to subject but even from assessment to assessment within subjects. A student's ability to insert an IV catheter in a patient is different from their ability to critique Thomas Hobbes' *Leviathan*, and so the assessment a student completes to show their capability in these areas is going to be different. What counts as valid for the former, is unlikely to count as valid for the later. What is needed is not a top down prescriptive assessment model to be given to instructors, but rather a conceptual toolkit by which instructors can understand what might count as appropriate assessments for their students. The distinction between structural and discursive changes, we argue, is a crucial first tool.

Conclusion

The emergence of GenAI has forced a fundamental reconsideration of assessment design in higher education. While many institutions have responded with frameworks that attempt to

guide AI use through increasingly sophisticated systems of rules, permissions, and declarations, these approaches fundamentally misunderstand the nature of the challenge. The distinction between structural and discursive changes reveals why current approaches, however well suited to other aims, ultimately prove insufficient to ensure assessment validity.

These frameworks remain powerless to prevent AI use when they rely solely on student compliance. They say much but change little. They direct behaviour they cannot monitor. They prohibit actions they cannot detect. In other words, when it comes to appropriate assessment change for a time of AI, talk is cheap.

The time invested in developing and implementing these discursive approaches is time that could otherwise be used to consider structural changes that will actually work to ensure assessment validity as well as the veracity and reputation of our degrees. When assessment validity hinges on student compliance with unenforceable rules rather than on inherent assessment design, we build educational systems on foundations of sand. Long term solutions require fundamentally rethinking how assessments are structured rather than how they are explained.

Looking forward, the challenge of AI in assessment will only intensify as AI capabilities continue to advance. The path forward through this increasingly challenging terrain lies not in more sophisticated rules about AI use, but in fundamentally redesigning how we structure assessments to demonstrate student capability. This will require significant effort and creativity from educators but has the advantage of allowing for genuine solutions to maintaining assessment validity in an AI-enabled world. These must be solutions based not on what we say, but on what we do.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Thomas Corbin http://orcid.org/0000-0001-8750-5711
Phillip Dawson http://orcid.org/0000-0002-4513-8287
Danny Liu http://orcid.org/0000-0002-6618-5576

References

Ardito, C. G. 2024. "Generative AI Detection in Higher Education Assessments." *New Directions for Teaching and Learning* 1–18. doi:10.1002/tl.20624.

Bearman, M., P. Dawson, S. Bennett, M. Hall, E. Molloy, D. Boud, and G. Joughin. 2017. "How University Teachers Design Assessments: A Cross-Disciplinary Study." *Higher Education* 74 (1): 49–64. doi:10.1007/s10734-016-0027-7.

Boud, D. 2000. "Sustainable Assessment: Rethinking Assessment for the Learning Society." *Studies in Continuing Education* 22 (2): 151–167. doi:10.1080/713695728.

Carless, D. 2017. "Scaling Up Assessment for Learning: Progress and Prospects." In *Scaling Up Assessment for Learning in Higher Education*, edited by D. Carless, S. M. Bridges, C. K. Y. Chan, and R. Glofcheski, 3–17. Singapore: Springer Singapore.

Corbin, T., P. Dawson, K. Nicola-Richmond, and H. Partridge. 2025. "Where's the Line? It's an Absurd Line': Towards a Framework for Acceptable Uses of AI in Assessment." *Assessment & Evaluation in Higher Education* 1–13. doi:10.1080/02602938.2025.2456207.

Dalalah, D., and O. M. Dalalah. 2023. "The False Positives and False Negatives of Generative AI Detection Tools in Education and Academic Research: The Case of ChatGPT." *The International Journal of Management Education* 21 (2): 100822. doi:10.1016/j.ijme.2023.100822.

Dawson, P., M. Bearman, M. Dollinger, and D. Boud. 2024. "Validity Matters More than Cheating." *Assessment & Evaluation in Higher Education* 49 (7): 1005–1016. doi:10.1080/02602938.2024.2386662.

Gonsalves, C. 2024. "Addressing Student Non-Compliance in AI Use Declarations: Implications for Academic Integrity and Assessment in Higher Education." *Assessment & Evaluation in Higher Education:* 1–15. doi:10.1080/02602938.2024.2415654.

Jiang, Z., J. Zhang, and N. Z. Gong. 2023. "Evading Watermark Based Detection of AI-Generated Content." In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 1168–1181. New York, NY: Association for Computing Machinery.

King's College London. n.d. "Generative AI: Guidance for Students." https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance/student-guidance

Lancaster University. n.d. "University Position on Artificial Intelligence." https://portal.lancaster.ac.uk/ask/administration/policies-regulations/university-position-on-artificial-intelligence/#d.en.599816

Liu, D., and A. Bridgeman. 2023. "What to Do about Assessments If we Can't out-Design or out-Run AI? Teaching@Sydney." https://educational-innovation.sydney.edu.au/teaching%40sydney/what-to-do-about-assessments-if-we-cant-out-design-or-out-run-ai/

Macquarie University. n.d. "Generative AI Assessment Checklist." https://ishare.mq.edu.au/prod/file/1dee0980-5ff7-4c07-9557-e95d6b26f051/1/GenAI-AssessmentChecklist.docx

Monash University. n.d. "Acknowledging the Use of Generative Artificial Intelligence." https://www.monash.edu/student-academic-success/build-digital-capabilities/create-online/acknowledging-the-use-of-generative-artificial-intelligence

Nikolic, Sasha, Isabelle Wentworth, Lynn Sheridan, Simon Moss, Elisabeth Duursma, Rachel A. Jones, Montserrat Ros, and Rebekkah Middleton. 2024. "A Systematic Literature Review of Attitudes, Intentions and Behaviours of Teaching Academics Pertaining to AI and Generative AI (GenAI) in Higher Education: An Analysis of GenAI Adoption Using the UTAUT Framework." *Australasian Journal of Educational Technology* 40 (6): 56–75. doi:10.14742/ajet.9643.

Perkins, M., L. Furze, J. Roe, and J. MacVaugh. 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 (06): 49–66. doi:10.53761/q3azde36.

Perkins, M., J. Roe, and L. Furze. 2024. "The AI Assessment Scale Revisited: A Framework for Educational Assessment." *arXiv Preprint* arXiv:2412.09029.

Perkins, M., J. Roe, B. H. Vu, D. Postma, D. Hickerson, J. McGaughran, and H. Q. Khuat. 2024b. "GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education." arXiv *P*reprint arXiv:2403.19148.

Princeton University. n.d. "Disclosing the Use of AI at Princeton University." https://libguides.princeton.edu/generativeAI/disclosure

St-Onge, C., M. Young, K. W. Eva, and B. Hodges. 2017. "Validity: One Word with a Plurality of Meanings." *Advances in Health Sciences Education: Theory and Practice* 22 (4): 853–867. doi:10.1007/s10459-016-9716-3.

Tertiary Education Quality and Standards Agency. 2024. "Gen AI Strategies for Australian Higher Education: Emerging Practice." https://www.teqsa.gov.au/sites/default/files/2024-11/Gen-AI-strategies-emerging-practice-toolkit.pdf

University College Dublin, College of Arts and Humanities. n.d. "Traffic Light System." https://www.ucd.ie/artshumanities/study/aifutures/trafficlightsystem/

University of Bath. 2024. "ABC Assessment Categorisations Guidance and Template Text." https://teachinghub.bath.ac.uk/guide/generative-ai-abc-assessment-categorisations/

University of Cambridge. n.d. "Template Declaration of the Use of Generative Artificial Intelligence." https://www.cshss.cam.ac.uk/education/generative-artificial-intelligence-ai-and-scholarship/template-declaration-use-generative

University of California, San Francisco. n.d. "Policy on Use of Artificial Intelligence (AI) in Assessments and Coursework." https://pharm.ucsf.edu/current/policies/ai

University of Leeds. n.d. "Categories of Assessments." https://generative-ai.leeds.ac.uk/ai-and-assessments/categories-of-assessments/

University of Melbourne. n.d. "Acknowledging AI Tools and Technologies." https://students.unimelb.edu.au/academic-skills/resources/academic-integrity/acknowledging-AI-tools-and-technologies

University of Reading. 2024. "Generative AI Tools and Assessment (Version 1.3)." Centre for Quality Support and Development. https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/ade/genai-tools-and-assessment.pdf

University of Sydney. 2024. "University of Sydney's AI Assessment Policy: Protecting Integrity and Empowering Students." https://www.sydney.edu.au/news-opinion/news/2024/11/27/university-of-sydney-ai-assessment-policy.html