Eesti Haridusteaduste Ajakiri, nr 14(2), 2026, 41–68
https://doi.org/10.12697/eha.2026.14.2.02b

Seven lessons for AI-robust assessment design in higher education

Izaak Dekker[1]

Centre for Applied Research of Education,
Amsterdam University of Applied Sciences

Abstract

The widespread adoption of AI by students in higher education has created urgent challenges for assessment design. As 89% of students use AI in their studies and nearly 80% use it for assessments, institutions are struggling to balance preventing misconduct with preparing students for responsible AI use. Existing reviews address parts of this challenge, but none comprehensively synthesize practical lessons for assessment design. This narrative review synthesized findings from recent empirical articles, previous reviews, and other recent studies into lessons that offer practical guidance to higher education professionals. Seven lessons are identified for AI-robust assessment design: (1) preventing cheating must not come at unacceptable costs to validity; (2) discursive solutions (e.g., AI use policies) must be combined with structural redesign; (3) a two-lane approach distinguishing independent from AI-augmented performance is advisable; (4) assessment design should protect the intrinsic cognitive load while allowing offloading of the extraneous load; (5) evaluative judgment has become a more central skill that requires extensive domain knowledge; (6) a diverse portfolio of assessment methods (the Swiss-cheese model) reduces vulnerability to both cheating and validity bias; and (7) AI-robust assessment requires sustained, integrated teacher professional development. Assessment design in times of AI is a wicked problem without quick fixes. The seven lessons offer a practical framework for institutions seeking to maintain validity, support learning, and uphold academic integrity in an AI-rich environment.

Keywords: assessment; artificial intelligence; higher education; assessment design

[1] Centre for Applied Research of Education, Amsterdam University of Applied Sciences, Wibautstraat 2 1092 GM, Amsterdam, The Netherlands, i.dekker@hva.nl.

Introduction

A large percentage of students use AI[2] in their studies in general (89% according to Chung et al., 2024) and specifically to help them with assessments (79.5% according to Gruenhagen et al., 2024). Many teachers and institutions encourage the responsible use of AI by students for the potential benefits it may bring (An et al., 2025; Luo et al., 2025; Moorhouse et al., 2023). Using AI may improve writing quality and the performance of students (Fleckenstein et al., 2023), personalize the learning support (Molenaar, 2022), and support cognitive and emotional engagement (Guo & Wang, 2025). Simultaneously, however, an increasing amount of risks and negative effects are surfacing: use of AI may lead to cognitive offloading and cognitive dependency (Abbas et al., 2024; Bastani et al., 2025), increases in dishonesty and cheating (Ardito, 2025; Newton, 2025), decreases in equity (Akgun & Greenhow, 2022), breaches of privacy and other side-effects that are not monitored (Dekker et al., 2025).

The potential gains and risks make it desirable for higher education institutions to control if and how AI is used. This is not an easy endeavor, given that it is hard or even impossible to reliably detect the use of AI (Ardito, 2025; Fleckenstein et al., 2024; Kofinas et al., 2025; Perkins, 2023; Perkins et al., 2024b; Waltzer et al., 2024) and that students generally do not like to disclose if and how they used AI (Gonsalves, 2025). This raises the question of how assessment design can still be valid and reliable. How can assessment design validly assess whether students have done the learning work themselves (Ellis & Murdoch, 2024), while also preparing and teaching students to use AI responsibly?

Four previous reviews have answered parts of this question. Cui et al. (2025) conducted a scoping review to chart the assessment literacy that teachers should have, in order to (re)design their assessments. Their focus was limited to teacher skills rather than the broader consequences of assessment and curriculum, and their search included articles up to September 2024. Kangwa et al. (2025) explored the factors that promote responsible and effective use of GenAI tools with a systematic literature review. Their review offered one of the most comprehensive and recent overviews of the literature, but the focus was on AI used by students rather than assessment. Xia et al. (2024) wrote a scoping review on how generative artificial intelligence is transforming assessment in higher education. This review included empirical articles published before September 2023, which made it unsuitable for providing comprehensive advice. Newton and Jones (2025) offered pragmatic advice in their guide for educators about

[2] When I write about AI in this article, I refer mostly to one specific application of AI: large language models. In practice, the term "large language models" is less often used than terms such as "GenAI", "Chatbots", and "ChatGPT".

how to adjust assessments for artificial intelligence. Their advice focused on the medical domain and did not use the literature available. What is missing is a practical review that summarizes lessons from recent literature on how to both prevent AI-related fraud and support students in learning to use it responsibly.

Theory

Two conceptual matters that are central to the literature on assessment design in higher education are also of particular importance in relation to AI: how we define assessment and how we think about its validity and reliability.

Firstly, there is the definition of the concept of assessment. The emphasis of assessment can be on contributing to learning or on making a sound judgment about students. The former focus is often called *assessment for learning* and has given rise to several related concepts, such as formative feedback or formative assessment (e.g., Wiliam, 2011). Within this stream of scholarship, assessment provides feedback for both the student and the teacher and can occur in class, between peers, or among teachers and peers. Emphasis can also be placed on how assessment can best function as quality assurance (e.g., Miranda, 2025) in high-stakes evaluations. From this perspective, matters of reliability, such as assessment consistency and validity, and measuring what matters are important components of quality assurance.

Both emphases are impacted by AI and can influence how it is used or regulated. AI can, for example, be positioned as a tool for feedback, serving as a partner or replacement in feedback dyads of peers and teachers (Meyer et al., 2026). It is debatable whether AI mostly enhances or impedes these feedback interactions, but its influence is undeniable and demands that institutions and teachers collectively take a stance. AI also affects how well different types of assessment can still make accurate judgments about a student's capabilities when it is not possible to distinguish a student's authentic work from work that is completely AI-generated (e.g., Newton & Jones, 2025).

Secondly, the conceptualization of validity plays a central role in understanding how AI can and should influence assessment design. The concept of validity is generally used in three different ways in assessment literature: as a test characteristic, as an argument-based-evidentiary claim, or as a social imperative (St. Onge et al., 2017). As a test characteristic, it refers to the extent to which assessments measure what they are intended to measure. When using this definition, the emphasis lies on the test as a relatively context-independent tool: certain forms of assessments are valid for certain learning goals. As an arguments-based-evidentiary claim, validity stands for a process in which the validity of an assessment is studied and not seen as a static characteristic. When

validity is viewed from the social imperative perspective, attention is also paid to the social or societal consequences of tests. For example, the impact of using a certain type of assessment on the workplace for which students are being prepared. Validity issues tend to be structural or related to the focus of an assessment. From the first perspective, certain assessment types, such as essays, have become invalid after the introduction of ChatGPT. From the second perspective, it is important to monitor validity when AI is used for an exam. The third perspective warns about the equity risks of rewarding students with expensive AI higher grades and the societal risks of deskilling if AI-independence is not inferred from assessments. Reliability, as an important related concept, stands for the consistency of judgment or the level of noise and arbitrariness when absent (Kahneman et al., 2021). Not being able to detect AI use can undermine validity and reliability as the end product, used for assessment, can no longer be seen as an indicator of student's capabilities.

Cheating and fraud in the context of this article refer to knowingly using AI for assessment in a way that was prohibited by the institution or teacher. Unconsciously using AI in prohibited ways can be a communications issue that can be solved by designing and communicating clear policies on what is and is not allowed. When disclosed, consciously acting against the rules for responsible AI use can be an act of open rebellion against a policy, or an act of deception when it is not disclosed to avoid the consequences.

Conceptually, it is also important to define AI and the different potential it has. For the sake of clarity, I will use the term "AI" instead of "large language models" for readability. However, eventually, I will make a distinction between general-purpose AI and educational AI. General-purpose AI, such as Claude and Mistral, has no pedagogical restraints and can be used by students for basically anything. Educational AI, on the other hand, has built-in restrictions that are geared towards a learning goal and based on a learning mechanism (Dekker et al., 2026; Dekker & Hermans, 2026). This could be a restricted persona within a chatbot that does not provide suggestions or write drafts, but only provides certain feedback on written texts.

Based on the discussion above, it is important to ascertain which types of assessment design are more vulnerable to the described risks to validity and reliability. I am, therefore, introducing the term 'AI-robustness' to indicate the degree to which assessment design validity and reliability are negatively impacted by the availability of general-purpose AI. More robust methods for assessment design are less vulnerable. The purpose of the present review is to summarize and synergize the prescriptive lessons for improving the AI robustness of assessment design that can be deduced from the recent literature.

Method

Methods used in a narrative review are not necessarily systematic and often not disclosed. This creates the impression that there is a dichotomy between transparent and systematic approaches on the one hand and a completely opaque, anything-goes approach on the other. Ideally, however, methods should be tailored to the research question and context, which is why I tried to be as transparent as possible about the specific approach used in this narrative review.

The increase in publications on AI and assessment has coincided with an exponential increase in low-quality publications; therefore, I decided to use a more focused approach than traditional systematic literature review protocols, which start by including a large number of hits that may include duplicates and low-quality studies. In order to answer the main question, I searched for review studies on assessments in higher education published after 2023 to gain a broad understanding of the field. I also decided to zoom in on the leading outlet that published the most impactful papers on the topic based on the reference lists of the review studies, and to continue from there. This was achieved by firstly checking all 148 published articles (89 with assigned issues and 59 online) from the 'Assessment & Evaluation in Higher Education' journal of 2025. There are other publications on assessment or higher education, but this journal's focus is the best match for the research question and contains the seminal papers that keep appearing in reviews. After reading all the titles and abstracts of the 148 eligible articles, I included 10 articles that specifically addressed assessment and AI. Next, and in order to include more relevant studies outside of Assessment & Evaluation in Higher Education, I read all the titles and abstracts of the articles in the reference lists of the 10 included articles (the 'snowballing' technique), which led to the inclusion of an additional 14 articles that met the inclusion criteria. Finally, I checked which articles had cited the included primary studies (the 'reverse snowballing' technique), leading to the inclusion of 12 more articles and a final sample of 36 studies. All searches were conducted on the 30th and 31st December 2025. All reviews and included articles were studied and scanned for potential prescriptive lessons regarding assessment design in the age of AI. Clustering and synthesizing these lessons led to a total of 7 general directions for advice that are summarized in the results section. Table 1 in the appendix shows all the included studies, the lessons derived from them and how they were used in the final synthesis.

Lesson 1: Preventing cheating at the cost of validity

After students adopted AI for schoolwork, many exams became more vulnerable to cheating, and cheating indeed increased (Newton, 2025). One reaction was to trade written assessments and take-home exams for verbal assessments and in-class exams in order to be able to detect whether the student could independently show mastery of the learning goals (An et al., 2025; Moorhouse et al., 2023). While this may increase the degree to which one can determine what the student knows and can do without AI, it can also decrease the validity of the assessment (Dawson et al., 2024). Validity stands for the degree to which the awarded grades represent what they are meant to (Dawson et al., 2024). If all assignments become verbal, but the profession for which the degree prepares students requires them to be good writers or to come up with original ideas, then only verbally testing students in situ is a less valid approximation of the required capabilities. If the profession requires you to write well with access to resources and time, testing whether you can write on the spot without access to resources may not be valid. If the profession requires you to be original, we first need to define what original means (Luo, 2024). Methods of assessment are seldom perfectly aligned with the learning goal. Relying exclusively on one type of method will compound the biases of this method. Only testing verbally disadvantages students who are less verbally able, and vice versa for using written methods, leading to differences between domains. For example, reading and writing are more central to the humanities than to technical studies (Luo & Dawson, 2025). Overly focusing on preventing cheating can also impact on the relationship between teachers and students and transform the role of the teacher into being an "AI cop" (Luo et al., 2025; Luo & Dawson, 2025). It is, therefore, important to ensure that the prevention of cheating does not lead to unacceptable losses in the validity of the combined assessments that a course uses.

Lesson 2: Combining discursive and structural solutions

The introduction of AI in education created much confusion for both students and teachers about which types of AI use, if any, are allowed (Corbin et al., 2025a; 2025b; 2025c; Eaton et al., 2025). One way to increase this transparency was by creating taxonomies or 'traffic lights' for each separate course to communicate the type of AI use was authorized and conducive to the learning goals (e.g., Furze et al., 2024; Perkins et al., 2024a). In addition to making explicit what is allowed, many have argued that it would be good to provide explicit examples of the correct use of AI (DeLuca et al., 2025; Eaton et al., 2025; Gonsalves et al., 2025; Gruenhagen et al., 2024; Kangwa et al., 2025; Tight, 2024), to teach students about the limitations of AI (Tsao, 2025), and explicitly

teach students about integrity (Parson, 2025; Tight, 2024). Parson (2025) found some evidence that offering integrity courses decreased academic misconduct. These 'discursive' solutions (telling students what is good or allowed) can provide clarity and prevent misconduct based on misunderstandings. However, this does not sufficiently deter all students from misconduct (Newton, 2025). In the final analysis, discursive solutions still depend on the willingness of students to transparently declare their use of AI. Students do not tend to transparently report AI use, fearful of lower grades (Gonsalves, 2025). Currently, their fears are justified, given that teachers (especially in the humanities) often mark down students who use AI with judgments such as 'this shows less independence or laziness' (Liu & Dawson, 2025). The communication about the authorized use of AI is not enforceable and places the responsibility for integrity with students (Corbin et al., 2025b; Eaton et al., 2025). Corbin et al. (2025b) argued that we need more structural solutions in addition to the discursive traffic light methods.

In line with the metaphor of the traffic lights, one can imagine integrity modules as driving lessons. If we extend the metaphor, then structural solutions might be similar to speed bumps, roundabouts, and speed cameras. What suggestions for structural solutions does the literature offer? DeLuca et al. (2025), Moorhouse et al. (2023), and Xia et al. (2025) propose redesigning assessments that require creativity, collaboration, community, critical thinking, and contextuality. Corbin et al. (2025c; 2025d) and Xia et al. (2025) propose using more authentic forms of assessment and holistic forms of judgment that emphasize the process rather than the output. Authenticity of the assessment does not yet seem to impact the ability to safeguard or protect against AI misconduct (Kofinas et al., 2025), but holistic methods could provide a way to save the assessment forms, such as the essay in places where it still offers a valid test of the learning goals (Corbin et al., 2025d).

An important condition for this is that students share their interactions with AI transparently rather than just the final product. One way to gain more insight into learning processes might be to use in-class work. These provide excellent opportunities to monitor the process of learning and to organize feedback interactions with students about the process of using AI in different responsible ways. Class attendance may be conditional for making holistic, process-oriented and formative assessment strategies work. Currently, however, attendance is not a given. Studies show that longitudinally monitored attendance patterns in higher education record a steep decline in class attendance, from 43% pre-COVID-19 to a post-COVID-19 (and ChatGPT) average attendance rate of just 30.6% (Dekker & Theelen, 2025). Adding structural to discursive solutions will also require a different approach to attendance.

Lesson 3: The two-lane approach

Given that it is hard or even impossible to reliably detect AI use, Liu and Brigman (2023) proposed dividing assessments into 'two lanes'. Assessments in lane 1 ensure that students demonstrate mastery of knowledge or skills independently of AI, using verbal, skill-based, or simulation-based assessments or exams under controlled conditions. Assessments in lane 2 allow students to use AI at each step of the process, provided they can report their AI use transparently. This proposal aligns with the recommendations of Ardito (2025), Eaton et al. (2025), and Lodge et al. (2023), who all propose distinguishing between tests that aim to establish independent performance and enhance augmented, hybrid, or automated performance. Liu and Brigman (2023) argue that a middle ground is feasible, given that it is not possible to reliably detect AI use beyond the controlled settings of lane 1. Curtis (2025) states that learning to use AI responsibly, like any educational process, will require scaffolding, constraints, and guidance, which requires some type of 'middle lane'. Lane 2 should not be an open invitation to use AI without supervision. Liu and Brigman (2023) do not advise relying on the goodwill of the students, while Curtis (2025) asserts that most students will follow the instructions. In many generic university contexts, the seduction of using AI inappropriately is too much for many students (Newton, 2025). A middle ground to this discussion could be to use a two-lane approach and also to offer guidance and instruction in lane 2 in line with lesson 2. The discussion assumes that educators have no control over the technology. General-purpose AI is not pedagogical: a good teacher or parent will not do anything you ask. Educational AI, developed by a combination of teachers and programmers, could be a way in which the technology could become more pedagogical, allowing for enforceable authorizations of different types of use.

In summary, the lesson is that it is advisable to test independent performances in controlled conditions (lane 1). How often will depend on costs, validity, and time (see lesson 1). We will address this further in lesson 6. By having several lane 1 assessments in place, you increase the chances that students have achieved mastery, which is a prerequisite for responsible use of AI.

Lesson 4: Desirable and undesirable cognitive load and offloading

Ellis and Murdoch (2024) distinguish between the work, which is the output or artefact that a student makes or performs and the work of learning, which is the process through which this work is made. They argue that the work of learning cannot be outsourced because it would defeat the purpose of learning. It is like using a forklift truck to lift weights in a gym. Using AI can be conceptualized

as a form of cognitive offloading (Fan et al., 2025). The work of learning, or 'intrinsic' cognitive load, should not be offloaded (Molenaar, 2022; Perkins, 2023). A range of studies have shown that AI use increases performance (the work) but decreases the work of learning by offloading cognitive (Bastani et al., 2025) or metacognitive effort (Fan et al., 2025), at least if it is not constrained, controlled, or guided (Dekker & Hermans, 2026). This raises the question about what parts of a task can or should be offloaded? Findings from cognitive psychology show that effort itself should not be offloaded, given that the most effective forms of learning are effortful (de Bruijn et al., 2023). For this reason, effortful but effective forms of learning are called 'desirable difficulties'. Difficulties that are undesirable, however, can be considered types of extraneous load that may be offloaded. This goes back to the importance of validity stressed earlier. What is the type of effort, or the work of learning, that the student needs to experience in order to reach the learning goals? Models like the AIAS can support teachers in pinpointing this and relating it to the role that technology is allowed to play (Perkins et al., 2024a). Reading and writing are more intrinsic to the humanities and less so to technical studies (Luo & Dawson, 2025). It is more likely to be valid to allow students to offload parts of the writing process in an engineering assignment than in a writing course.

Lesson 5: Evaluative judgments

The popularity of the term Generative AI for most of the recent machine learning applications reflects its primary use: generating texts, images or content. If AI can generate it, then judging the quality of its output becomes more important (Bearman et al., 2024; Zhan & Yan, 2025). Tai et al. (2018) describe evaluative judgment as consisting of two components: a person must have an internal understanding of quality and apply this to a specific work. This is a disciplinary skill rather than a transferable general skill (Bearman et al., 2024; Tai et al., 2018). It requires domain knowledge, and can also be deliberately trained using self-assessment, peer-assessment and feedback, rubrics, and exemplars. In addition to assessing the output of AI, students and teachers can also evaluate the process of AI use and the judgment of student work by AI (Bearman et al., 2024). A novice is more likely to uncritically accept output, whereas users who are more knowledgeable about the particular subject or scrutinize it more closely are able to recognize problematic output and add iterations or adapt their AI use (Walton et al., 2025; Zhan & Yan, 2025). Students generally want the study work they produce to come from themselves (Bearman et al., 2025), but at the same time can misattribute output from AI as their own work (Walton et al., 2025). One could call this a meta-cognitive awareness

of AI dependency. This teaches us two things. Firstly, evaluative judgment has indeed become more important because of AI, but there are no shortcuts to obtaining it, as it requires both domain knowledge and deliberate practice with the known forms of effortful and effective assessment. Second, there is a high risk of the notorious 'Matthew' effect when it comes to evaluative judgment and AI: less knowledgeable students are more likely to uncritically accept output and thereby become less adept at evaluative judgment, while more knowledgeable students will spot weaknesses and engage in longer iterations that will increase their evaluative judgment (Dekker & Hermans, 2026). This process is further exacerbated by the personalized nature of chatbots, which adapt to users based on chat history by lowering or altering expectations and making new suggestions to close the loop. The prescriptive lesson to be drawn is that teachers play a central role in organizing the formative feedback processes to ensure that all students have the opportunity to improve their evaluative judgment. This will require providing general knowledge about AI literacy and the limitations of AI, but it will also still require the usual disciplinary deliberate practice.

Lesson 6: It takes a Swiss cheese and a village

In lesson 1, I mentioned the risks of relying on a single type of assessment. The literature on AI detection shows that there are few AI-proof assessment methods. The controlled lane 1 solutions can have a purpose, but turning everything into a controlled lane 1 assessment may hurt validity in several ways (lesson 1) because the biases in those assessment methods compound at the program level. The so-called Swiss-cheese model aims to tackle this exact problem (Rundle et al., 2020). Its proposed solution is to create a diverse portfolio of assessment methods to reduce vulnerability to the downsides of any one method. Each assessment type can be seen as a slice of Swiss cheese, which, like the famous Emmental, contains a few holes (air pockets). In the metaphor, the holes represent the downsides of that particular type of assessment. The model assumes that the holes or downsides are different from each other, thereby compensating for each other's weaknesses. This idea of 'spreading the risk' pertains to both the risk of cheating with AI and the risk of validity bias. This advice dates from before AI took off and has become even more important now.

The Swiss-cheese model requires a concerted team effort because its strength lies not in individual assessments but in the combination of assessments in the aggregate. In a similar vein, Nicola-Richmond et al. (2025) argue that it takes a whole team (a village) to make assessments more robust against AI. Assessment redesign in line with, for example, the lessons above, requires skills that are seldom combined in individuals but that might be found in complementary teams: assessment design, subject-matter expertise, program knowledge, industry

knowledge, and AI literacy. Another reason why it takes a whole team is that the controlled 'lane 1' assessments need to be strategically placed in the curriculum. It can be less valid but also extremely labor-intensive to turn each assessment into a controlled practicum or verbal assessment. Together, a team can decide at which points in the curriculum students need to show independent mastery of crucial skills before proceeding.

Lesson 7: The role of teacher professional development

AI poses several challenges for teachers and teaching teams. Assessment in times of AI has all the characteristics of a 'wicked' problem and cannot be 'solved' (Corbin et al., 2025a). Most teachers are motivated to redesign their assessments for AI (Bower et al., 2024), but it requires careful navigation and redesign from teachers and teaching teams who will need support to do this well (Corbin et al., 2025a; Corbin et al., 2025c; DeLuca et al., 2025) and currently often do not experience sufficient support (Nicola-Richmond et al., 2025). Applying the previous six lessons will require professional development to improve AI literacy, ethical dialogues about the affected values, and assessment design skills of teaching teams (Bower et al., 2024; Cui et al., 2025; DeLuca et al., 2025; Kangwa et al., 2025; Xia et al., 2024). Additionally, it will require teachers to model responsible AI use or avoid using it irresponsibly. Many students feel that there is a general unfairness in one-sidedly imposing AI norms on students when teachers themselves use AI to judge student work or create assessment questions (Luo et al., 2025). DeLuca et al. (2025) argue that professional development for responsible AI use and assessment redesign cannot be a simple add-on; it needs to be integrated into existing professional development structures. Many higher education institutions have 'centers for teaching and learning' or similar organizational units that can play a central role in helping to integrate the required professional development components into induction programs, teaching certificates, and fellowships. The broad range of professional development activities can use the breakthrough insights in the literature on effective professional development in education provided by the review, as captured in recent articles by Sims et al. (2025) and Visscher et al. (2025). Professional development should yield practical insights related to teaching and learning for educators and provide a clear definition of what "good" looks like using exemplars, and be clear about why this should be done. This will not be easy, given that AI is not neutral and is intertwined with professional and pedagogical values (Sporrong et al., 2025). If what counts as a "good" performance is clear, then teachers can compare their own competence with a "good" performance and plan how to improve it further. Effective professional development should also organize techniques for implementing insights

and help to embed changes in practice. This can be done by using classroom coaching and facilitating teacher cooperation (as a means of mutual support and knowledge exchange).

Discussion

The spike in scholarly attention for assessment design in the age of AI does not offer quick fixes to the wicked problem of AI assessment design (Corbin et al., 2025a). Publications from the past years, however, do offer the following seven lessons that can be used to make assessment more robust to the challenges that AI poses: 1) preventing cheating must not come at unacceptable costs to validity; (2) discursive solutions must be combined with structural redesign; (3) strategically organize some controlled assessments in which students cannot use AI; (4) assessment design should protect intrinsic cognitive load while allowing offloading of extraneous load; (5) evaluative judgment has become a more central skill that requires extensive domain knowledge; (6) a diverse portfolio of assessment methods (the Swiss-cheese model) reduces vulnerability to both cheating and validity bias; and (7) AI-robust assessment requires integrated evidence-informed teacher professional development.

Both assessment for learning and assessment as judgment are affected by the availability of AI to students. The potential for cheating and the lack of reliable detection can undermine both the reliability and validity of the assessment. As we learned in lessons 1 and 2, however, assessment for learning can be an important part of the solution: organizing more formative assessments during classes can reduce reliance on high-stakes assessment as the sole judgment and provide more information about the learning process. This means that assessment of learning approaches can no longer depend on portfolios or reports of learning processes but needs to make this visible during in-class tasks. Courses will need to strike the right balance between controlled assessments without AI at strategic moments and more ecologically valid assessments that are grounded in regular in-class assignments and feedback.

By synthesizing recent findings into seven lessons, this paper offers an overview that can be useful to practitioners who want to get up to speed on what we've learned thus far and researchers who want to build on the rapidly expanding and diffuse knowledge base. The seven lessons combined insights from previous reviews (Cui et al., 2025; Kangwa et al., 2025; Newton & Jones, 2025; Xia et al., 2024) with a focus on assessment and practical advice, and added new insights based on recent papers that were not yet included in those reviews (e.g., Bearman et al., 2025; Corbin & Walton, 2025; Luo & Dawson, 2025; Newton, 2025; Nicola-Richmond et al., 2025).

Future developments

The current lessons are based on existing literature, which, in turn, is based on versions of AI that are superseded by new developments as this article is being written. How will assessment design remain AI-robust as the technology keeps developing? Or, how future-robust are these lessons? There are three different likely types of directions that we need to consider: technological developments in general-purpose AI, developments in educational AI, and developments in cultural attitudes towards AI.

The first direction of development will mean ever more powerful general-purpose AI that can perform more tasks, in a better manner, and be more autonomous or 'agentic'. Given these developments, the same seven lessons will be even more urgent. More powerful AI is likely to impact the world for which higher education needs to prepare students, meaning that only banning AI from assessments will come at the cost of validity (lesson 1). It will also likely increase the importance of evaluative skills as more tasks are outsourced to AI (lesson 5).

Developments in educational AI can be illustrated with a recent example. Estonia made international headlines when its president presented the "AI Leap" (*TI-Hüpe*) project. The AI Leap is a public-private collaboration between companies, the Estonian government, schools, and researchers that makes a special version of ChatGPT available to Estonian students (Aru & Laak, 2025). This version of ChatGPT is programmed not to provide direct answers or drafts in response to student questions; instead, it is supposed to help students find the answers themselves in a Socratic fashion (Aru & Laak, 2025). Using such 'pedagogical constraints' could potentially counteract the user-friendly tendency to remove desirable difficulties for students (Dekker & Hermans, 2026). Such educational AI could preserve the intrinsic cognitive load and instead offer personalized formative feedback on demand (Laak & Aru, 2025). Fundamentally, however, it would not change the challenges to assessment: even in this situation, students will still also have access to general-purpose AI. Choosing educational AI over general-purpose AI will be up to the inner discipline of the student, unless teachers can completely control what technology students have access to in and outside of class.

The third direction involves cultural reactions to AI. Society, and in particular teachers, can come to embrace or reject AI across all domains of life. Currently, teachers hold balanced views on the potential benefits and risks of AI in and for education (Granström & Oppi, 2025; Kalmus & Nikiforova, 2024). In scenarios in which AI acceptance or rejection is unanimous, both are unlikely and relatively easy for policy: without AI, using a varied, valid assessment portfolio will still be desirable, and with widespread AI protective

measures described in the lessons, it will be necessary. However, it is more likely that we are moving towards a more polarized setting in which some completely reject AI use while others fully integrate it into their work processes. In this scenario, assessment design should both protect against the option of outsourcing intrinsic cognitive effort but also keep in mind that some teachers and students will not use AI. This affects the degree to which in-class lessons could be used to model and practice with responsible AI use, and it makes for an uneven playing field during uncontrolled assessments ('lane 2'). The seven lessons do not provide a solution to this predicament.

Limitations and future directions

This review synthesized practical lessons for assessment design in times of AI in higher education. It used a narrative approach to offer a concise reading experience for a broad audience of higher-education teachers and researchers. The downside of not using a systematic approach is that this review does not offer a transparent and precise focus that is replicable. This was a conscious trade-off, given the availability of many systematic and scoping reviews that were also included in this article.

Furthermore, by focusing on practical advice, it did not question some of the more fundamental normative questions about the desirability of (different types of) AI technology itself (e.g., Wieczorek, 2025). Nor did it synthesize or tackle the empirical question of whether the gains outweigh the negative effects of AI on learning (Dekker et al., 2025; 2026). These questions were beyond the scope of this paper but deserve at least as much attention and will influence questions of assessment design. Controlled 'lane 1' solutions (Liu & Brigman, 2023) will receive more emphasis if the academic community decides that AI in the form of large-language models offered by commercial companies is unethical. Think of, for example, the stolen data on which they are trained, privacy risks, inherent unreliability, and their ecological footprint (Dekker et al., 2025). Rejection of general-purpose AI could also eventually lead to its technical adaptation for educational purposes, either offered by companies, home-built by public organizations, or private-public consortia, which could impact the enforceability of authorized use. Empirical breakthroughs in our knowledge about the added value of AI for learning will and should equally influence the degree to which educational organizations want to ban or stimulate different types of use.

Future research should prioritize experimental studies that track the actual effects of assessment redesign on student learning and academic integrity, rather than relying on cross-sectional surveys of student or teacher attitudes. There is also a pressing need for research on effective educational AI (as distinct from

general-purpose AI) and its potential to create more enforceable, pedagogically grounded conditions for AI use in assessment. Initiatives such as the AI Leap project in Estonia could offer important contributions towards these developments.

Conclusion

The seven lessons presented in this review collectively point to a fundamental tension that higher education institutions must navigate: the need to maintain the integrity and validity of assessment as a trustworthy credential, while simultaneously preparing students for a professional world in which AI is a routine tool. This tension is not new. Similar debates accompanied the introduction of calculators, spell-checkers, and the internet, but the speed and breadth of AI adoption, combined with the difficulty of detecting its use, make the current moment qualitatively different (Newton, 2025; Perkins, 2023). The lessons in this review suggest that institutions cannot resolve this tension by banning AI use entirely or by permitting its use without structural measures and guidance (Curtis, 2025; Dawson et al., 2024).

What emerges instead is a picture of assessment design as an ongoing, collaborative, and values-laden practice rather than a technical problem with a stable solution. This aligns with broader calls in the assessment literature for a shift away from assessment as measurement towards assessment as a learning and developmental process (Corbin & Walton, 2025). AI intensifies this shift: when the production of written output can be largely automated, the educational value of assessment must increasingly be located in the learning process. Additionally, domain knowledge becomes increasingly important due to its instrumental role in evaluating AI-generated content. Perhaps surprisingly, this means that making education AI-robust will also entail finding ways to revalue and strengthen two classic core tenets of education: class attendance and engagement, and domain knowledge accumulation.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, *21*(1), 10. https://doi.org/10.1186/s41239-024-00444-7

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, *22*(10), https://doi.org/10.1186/s41239-025-00507-3

Ardito, C. G. (2025). Generative AI detection in higher education assessments. *New Directions for Teaching and Learning*, *182*, 11–28. https://doi.org/10.1002/tl.20624

Aru, J., & Laak, K. J. (2025). Developing an AI-based general personal tutor for education, *Trends in Cognitive Sciences*, *29*(11), 957–960. https://doi.org/10.1016/j.tics.2025.09.010

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Bearman, M., Fawns, T., Corbin, T., Henderson, M., Liang, Y., Oberg, G., Walton, J., & Matthews, K. E. (2025). Time, emotions and moral judgements: How university students position GenAI within their study. *Higher Education Research & Development*, 1–15. https://doi.org/10.1080/07294360.2025.2580616

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

Bower, M., Torrington, J., Lai, J.W.M., Petocz, P. & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technology*, *29*, 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Chung, J., Henderson, M., Pepperell, N., Slade, C., & Liang, Y. (2024). Student perspectives on AI in higher education: Student survey. *Student Perspectives on AI in Higher Education Project*. https://doi.org/10.26180/27915930

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025a). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 1–17. 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*, 1–11. https://doi.org/10.1080/02602938.2025.2503964

Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025c). '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. https://doi.org/10.1080/02602938.2025.2456207

Corbin, T., Walton, J., Bannister, P., & Deranty, J. P. (2025d). On the essay in a time of GenAI. *Educational Philosophy and Theory*, 1–13. https://doi.org/10.1080/00131857.2025.2572802

Corbin, T. A., & Walton, J. (2025). The missing story of GenAI summarisers: a critical research agenda. *Higher Education Research & Development*, 1–14. https://doi.org/10.1080/07294360.2025.2486185

Cui, Y., Meng, Y., & Tang, L. (2025). Reconsidering teacher assessment literacy in GenAI-enhanced environments: A scoping review. *Teaching and Teacher Education*, *165*, 105163. https://doi.org/10.1016/j.tate.2025.105163

Curtis, G. J. (2025). The two-lane road to hell is paved with good intentions: why an all-or-none approach to generative AI, integrity, and assessment is insupportable. *Higher Education Research & Development*, 1–8. https://doi.org/10.1080/07294360.2025.2476516

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

de Bruin, A. B., Biwer, F., Hui, L., Onan, E., David, L., & Wiradhany, W. (2023). Worth the effort: The start and stick to desirable difficulties (S2D2) framework. *Educational Psychology Review*, *35*(2), 41. https://doi.org/10.1007/s10648-023-09766-w

Dekker. I., Agarwal, B., Bredeweg, B., van de Poel (2026). Usable procedures for responsible experiments with artificial intelligence in education. *AI and Ethics*, https://doi.org/10.1007/s43681-026-01035-0

Dekker, I., Bredeweg, B., te Winkel, W., & van de Poel, I. (2025). Ethical procedures for responsible experimental evaluation of AI-based education interventions. *AI and Ethics*, *5*(3), 2977–2986. https://doi.org/10.1007/s43681-024-00621-4

Dekker, I. & Hermans, F. (2026). Why the user-friendliness of AI undermines education. Commentary on Wieczorek 2025. *Philosophy & Technology*. *39*, 82. https://doi.org/10.1007/s13347-026-01093-6

Dekker, I. & Theelen, H. (2025). The vanishing classroom: Attendance trends in higher education. *OSF*. https://doi.org/10.35542/osf.io/bmnhp_v2

DeLuca, C., Volante, L., & Holden, M. (2025). The AI3 model: Future directions for artificial intelligence, assessment innovation, and academic integrity. *Educational Researcher*, 0013189X251385537. https://doi.org/10.3102/0013189X251385537

Eaton, S. E., Moya Figueroa, B. A., McDermott, B., Kumar, R., Brennan, R., & Wiens, J. (2025). What should we be assessing exactly? Higher education staff narratives on gen AI integration of assessment in a postplagiarism era. *Assessment & Evaluation in Higher Education*, 1–20. https://doi.org/10.1080/02602938.2025.2587246

Ellis, C., & Murdoch, K. (2024). The educational integrity enforcement pyramid: a new framework for challenging and responding to student cheating. *Assessment & Evaluation in Higher Education*, *49*(7), 924–934. https://doi.org/10.1080/02602938.2024.2329167

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, *6*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology*, *40*(4), 38–55. https://doi.org/10.14742/ajet.9434

Gonsalves, C. (2025). Addressing student non-compliance in AI use declarations: implications for academic integrity and assessment in higher education. *Assessment & Evaluation in Higher Education*, *50*(4), 592–606. https://doi.org/10.1080/02602938.2024.2415654

Granström, M., & Oppi, P. (2025). Assessing teachers' readiness and perceived usefulness of AI in education: an Estonian perspective. *Frontiers in Education*, *10*, 1622240. https://doi.org/10.3389/feduc.2025.1622240

Gruenhagen, J., P.M. Sinclair, J. Carroll, P. Baker, A. Wilson, & D. Demant. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence*, *7*, 100273. https://doi.org/10.1016/j.caeai.2024.100273

Guo, Y., & Wang, Y. (2025). Exploring the effects of artificial intelligence application on EFL students' academic engagement and emotional experiences: A Mixed-Methods study. *European Journal of Education*, *60*(1), e12812. https://doi.org/10.1111/ejed.12812

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Hachette UK.

Kalmus, J.E., & Nikiforova, A. (2024). Generative AI adoption in higher education: Exploring educator resistance in Estonian universities. *Proceedings of the 2024 Pre-ICIS SIGDSA Symposium*. 13. https://aisel.aisnet.org/sigdsa2024/13

Kangwa, D., Msafiri, M. M., & Fute, A. (2025). Exploring the factors that promote a balance between academic integrity and the effective use of GenAI tools in higher education: A systematic review. *Journal of Computer Assisted Learning*, *41*(5), e70109. https://doi.org/10.1111/jcal.70109

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*, 2522–2549. https://doi.org/10.1111/bjet.13585

Laak, K. J., & Aru, J. (2025). AI and personalized learning, *Educational Technology & Society*, *28*(4), 133–150, https://doi.org/10.30191/ETS.202510_28(4).RP08

Liu, D., & Bridgeman, A. (2023). *What to do about assessments if we can't out-design or out-run AI?* https://educational-innovation.sydney.edu.au/teaching@sydney/what-to-do-about-assessments- if-we-cant-out-design-or-out-run-ai/

Lodge, J., Howard, S., Bearman, M., & Dawson, P. (2023). Assessment reform for the age of artificial intelligence. https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf

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

Luo, J., & Dawson, P. (2025). Exploring value judgements in grading: will teachers mark down student work assisted by GenAI, and should they? *Studies in Higher Education*, 1–15. https://doi.org/10.1080/03075079.2025.2552825

Luo, J., Keung, C. P. C., & Tang, H. Hang H. (2025). Assessment as a dilemmatic space in the GenAI age: mapping and unpacking university teachers' conflicting priorities in assessment. *Assessment & Evaluation in Higher Education*, *50*(4), 607–621. https://doi.org/10.1080/02602938.2024.2444890

Meyer, J., Köller, O., Jansen, T., Fleckenstein, J., Asher, M. W., Bichler, S., ... & Winstone, N. (2026). The future of feedback: How can AI help transform feedback to be more engaging, effective, and scalable? *arXiv*, *2603.12463*. https://doi.org/10.48550/arXiv.2603.12463

Miranda, F. J. (2025). Accreditation and quality assurance in higher education institutions: a systematic literature review and a research agenda. *Quality in Higher Education*, *31*(3), 304–320. https://doi.org/10.1080/13538322.2025.2553983

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. https://doi.org/10.1016/j.caeo.2023.100151

Newton, P. M. (2025). How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment. *Assessment & Evaluation in Higher Education*, *50*(8), 1332–1343. https://doi.org/10.1080/02602938.2025.2511794

Newton, P. M., & Jones, S. (2025). Education and training assessment and artificial intelligence. A pragmatic guide for educators. *British Journal of Biomedical Science*, *81*, 14049. https://doi.org/10.3389/bjbs.2024.14049

Nicola-Richmond, K., Dawson, P., Partridge, H., & Macfarlane, S. (2025). It takes a village... Program-wide approaches to redesigning assessment in a time of generative artificial intelligence (GenAI). *Journal of University Teaching & Learning Practice*, *22*(7). https://doi.org/10.53761/zpp2ja61

Parson, V. (2025). Beyond penalties: a proactive approach to developing student understanding of academic integrity. *Assessment & Evaluation in Higher Education*, 1–16. https://doi.org/10.1080/02602938.2025.2580621

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/10.53761/1.20.02.07

Perkins, M., L. Furze, J. Roe, & 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. https://doi.org/10.53761/q3azde36

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024b). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, *22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6

Rundle, K., Curtis, G., & Clare, J. (2020). Why students choose not to cheat. In *A research agenda for academic integrity* (pp. 100–111). Edward Elgar Publishing.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., ... & Anders, J. (2025). Effective teacher professional development: New theory and a meta-analytic test. *Review of Educational Research*, *95*(2), 213–254. https://doi.org/10.3102/00346543231217480

Sporrong, E., C. McGrath, & T. Cerratto Pargman. (2025). Situating AI in assessment—An exploration of university teachers' valuing practices. *AI and Ethics*, *5*(3), 2381–2394. https://doi.org/10.1007/s43681-024-00558-8

St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: one word with a plurality of meanings. *Advances in Health Sciences Education*, *22*(4), 853–867. https://doi.org/10.1007/s10459-016-9716-3

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

Tight, M. (2024). Challenging cheating in higher education: A review of research and practice. *Assessment & Evaluation in Higher Education*, *49*(7), 911–923. https://doi.org/10.1080/02602938.2023.2300104

Tsao, J. (2025). Trajectories of AI policy in higher education: Interpretations, discourses, and enactments of students and teachers. *Computers and Education: Artificial Intelligence*, 100496. https://doi.org/10.1016/j.caeai.2025.100496

Visscher, A. J., Dmoshinskaia, N., Pellegrini, M., & Rey-Naizaque, A. (2025). (When) do teacher professional development interventions improve student Achievement? A meta-analysis of 128 high-quality studies. *Educational Research Review*, *49*, 100742. https://doi.org/10.1016/j.edurev.2025.100742

Walton, J., Bearman, M., Crawford, N., Tai, J., & Boud, D. (2025). How university students work on assessment tasks with generative artificial intelligence: Matters of judgement. *Assessment & Evaluation in Higher Education*, *50*, 1–17. https://doi.org/10.1080/02602938.2025.2570328

Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity*, *20*(1), 11. https://doi.org/10.1007/s40979-024-00158-3

Wieczorek, M. (2025). Why AI will not democratize education: a critical pragmatist perspective. *Philosophy & Technology*, *38*(2), 53. https://doi.org/10.1007/s13347-025-00883-8

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), 40. https://doi.org/10.1186/s41239-024-00468-z

Zhan, Y., & Yan, Z. (2026). Students' engagement with ChatGPT feedback: implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *51*(5), 821–834. https://doi.org/10.1080/02602938.2025.2471821

Appendix

Table 1

Study	Findings	Interpretation and application
An et al. (2025)	Most higher education institutions have faculty policies for AI that prescribe course-specific policies.	Lessons 2 and 7: Policies currently expect teachers to come up with course-specific policies that should prevent fraud and stimulate responsible use. This is a big ask.
Ardito (2025)	AI-detection tools are misaligned and unreliable; Use reliable assessments to test core functions students need to do without LLMs and increase standards for performance with LLMs (similar to calculators).	Lesson 3: Similar to the introduction of calculators, it may be smart to reliably test whether fundamental skills have been learned without dependency and then allow use for higher performance.
Bearman et al. (2025)	Students prioritize their own moral compass over institutional messaging and generally think that their educational work should come from and by owned by themselves.	Lesson 5: Effects of institutional messaging may be limited; students desire to ensure the work comes from themselves.
Bearman et al. (2024)	We should help students to develop capabilities to identify and calibrate the quality of work through existing formative assessment strategies; Students can misattribute work done by AI to themselves.	Lesson 5: Responsible use of AI hinges on the capabilities of the users to assess the quality of the output, we should emphasize strategies that improve these capabilities.
Bower et al. (2024)	Teachers with more knowledge of AI expected a higher impact on practice. They are most motivated to change their teaching and assessment practices to increase the performance expectancy of their students and themselves.	Lesson 7: The need for professional development of teachers with regard to AI could tap into their motivation for changing assessment practices to increase performance expectancy.
Cui et al. (2025)	There is a need for targeted professional development of teachers. The teacher assessment literacy (AL) in GenAI-enhanced environments (TALiGAI) model can help to tailor support and PD.	Lesson 7: Teacher professional development should not treat teachers as blank slates when it comes to AI literacy, and also take emotional and contextual factors into account.

Study	Findings	Interpretation and application
Corbin et al. (2025a)	AI and assessment have all the character traits of a 'wicked problem'. This means that there are no clear solutions. Institutions should recognize this and offer leeway and support to teachers who have to navigate solutions in their practices.	Lessons 2 and 7: There is no clear policy guidance to fix the issue; We need to consider what we are asking from teachers and offer support and PD that can help them deal with this wicked problem.
Corbin et al. (2025b)	In the first instance, many institutions have resorted to discursive traffic light solutions. These one-sidedly put the responsibility for integrity on the students and do not suffice; we need more structural solutions.	Lesson 2: Offer structural solutions in addition to discursive ones.
Corbin et al. (2025c)	Assessment policies for AI should address three critical dimensions: the feasibility of enforcement, the preservation of authentic learning, and the emotional well-being of teachers and students.	Lesson 1: Ensure that assessment policies do not undermine authenticity and validity. Lessons 2 and 7: Be mindful of the wicked and uncertain nature of the problem and the burden that navigating this entails for teachers and students.
Corbin et al. (2025d)	The validity of the essay for assessment is compromised. Assessing it more holistically instead of grading the product itself may offer a solution.	Lesson 2: Example of how a more holistic assessment can function as a structural solution.
Curtis (2025)	The two-lane approach (no AI or unconstrained use of AI) does not allow for middle-ground assessments where GenAI use is permitted within limits and assessment security is partial.	Lesson 3: The two-lane approach could be complemented with a middle lane for constrained AI use.
Dawson et al. (2024)	Attempts to prevent students from cheating during assessment may undermine its validity, in turn undermining its purpose and making it less inclusive.	Lesson 1: When making AI more robust, be mindful of the 'price' that is paid. Switching to only verbal assessments could be less valid if the profession does not require only verbal skills and disadvantages verbally weaker students.

Study	Findings	Interpretation and application
Deluca et al. (2025)	Students and teachers need AI literacy to understand its limits. AI forces structural changes in assessment policy from classic written forms to innovative ones that center on creativity, collaboration and community. The AI^3 model provides suggestions for communication to students, assessment design, and teacher professional development.	Lesson 2: Be clear about authorized use and about transparent reporting of AI use, and cultivate an understanding of originality. Teachers should model this behavior. Lesson 7: Provide resources and ongoing professional development for teachers to navigate and improve their assessment methods in innovative ways that focus on creativity, collaboration and community.
Eaton et al. (2025)	Assessment may have to shift from the output (essay) to the process, and it may be unethical not to teach students how to use it critically, understand its flaws, and evaluate its output.	Lesson 2: We need to teach students how to use it with integrity or when not to use it. Lesson 5: We need to help students evaluate output and use a more holistic process evaluation to assess learning.
Ellis & Murdoch (2024)	Cheating-detection and prevention strategies should differentiate among types of student behavior (tiered policy). Currently, most universities have a detection problem and only have a bottom tier (integrity instructions) and a top tier (expulsion) in place, with little in between that can create 'downward pressure'.	Lesson 4: The learning work cannot be outsourced. When this is done, it undermines the value of the diploma. Universities need to find ways to monitor and detect whether this is occurring and have a range of tiered measures and consequences (from honor or educational-cooperative to more punitive) in place that deter cheating.
Furze et al. (2024)	The AIAS model can help teachers establish what AI usage should be authorized based on the learning goals and help to clearly communicate AI use in each course. Applying it in a university led to fewer notifications of academic fraud and stimulated teachers to redesign assessment.	Lesson 2: Clarity about what is and is not allowed can help students and teachers if combined with the redesign of assessment. Lesson 3: Several lanes can be beneficial for differentiating in how AI use is authorized per condition.
Gonsalves (2025)	Many studies recommend policies that require students to transparently describe how they used AI. These studies found that students are generally unwilling to declare this out of fear of potential consequences.	Lesson 2: Simply asking students to use AI responsibly and declaring, if and how they used it, is not enough.

Study	Findings	Interpretation and application
Gruenhagen et al. (2024)	80% of students use AI to assist them with assessment. There is a large variance in what students consider acceptable.	Lesson 2: Teachers should provide transparency about what is allowed, but also make structural changes to the assessment.
Kangwa et al. (2025)	Structured support and academic training, technical scaffolding and perceived usefulness enable ethical AI use. Transparent, adaptive, and discipline-responsive policies correlate with less academic misconduct.	Lessons 2 and 7: Structured support and academic training, technical scaffolding, and perceived usefulness enable ethical AI use.
Luo (2024)	Higher education institutions' AI policies assume a problematic notion of originality that cannot be clearly delineated in practice. They also focus their frame mostly on catching plagiarizing students rather than engaging students to do the right thing.	Lesson 1: There are trade-offs and risks involved with focusing on stopping or preventing cheating. Do we even have a clear notion of originality that we can uphold?
Luo & Dawson (2025)	Grading is influenced by person-oriented values about integrity (hiding AI use) or laziness in relation to AI (outsourcing or not bothering to let AI check mistakes), capability-oriented values (AI-[in]dependent performance shows higher capability), and justice-oriented values (fairness and beneficence)	Lesson 1: Teacher values can interfere with validity. How AI use is considered when grading can change the relationship between students and teachers. Lesson 2: Clear guidelines can reduce implicit and non-transparent value judgments that reduce validity Lesson 7: Teacher PD could help teachers become aware of how values influence grading.
Luo et al. (2025)	Teachers identify AI-enabled cheating as their key priority to address in assessment. This influences their role as teachers, potentially turning them into "AI cops". Dilemma's surrounding their own use of GenAI for grading and feedback provision are underdiscussed.	Lesson 1: Focusing on preventing AI cheating might hurt the teacher student relationship. Lesson 7: Teachers might not yet be sufficiently aware of the dilemma's concerning their own use of GenAI.
Moorhouse et al. (2023)	As a teacher, first, run your assessment task through an LLM to see how it will be generated. Train teachers in LLM assessment literacy.	Lesson 3: In lane 2, emphasize critical thinking, creativity and contextuality. Focus on process and in-class feedback.

Study	Findings	Interpretation and application
Newton (2025)	Many courses still use assessment forms, such as essays, that are vulnerable to cheating. The overwhelming majority of students use AI to help with these assessments, and 22% self-declare that they use AI to cheat (contrary to the policy). Examples of cheating range from support in paraphrasing to writing the essay for them.	Lesson 2: Many students are aware of the policies regarding AI and assessment, and a substantial number do not adhere to them. Lesson 7: A lot of assessments have not yet been redesigned or adjusted and are prone to academic misconduct. Teachers will need support with doing this.
Newton & Jones (2025)	We cannot skip testing basic knowledge because higher-order thinking skills depend on it. Some forms of assessment need to be supervised or changed in order to be valid. Validity should guide choices. Integrity and possessing the skills are integral to becoming a responsible medical professional.	Lesson 3: We need lane 1 to test basic knowledge before it can be applied. Multiple-choice questions can still be used under supervised conditions, but the essay is no longer viable, at least in medical education, where it is less valid to write about doing something than actually doing it.
Nicola-Richmond et al. (2025)	Making assessment robust to AI requires program-wide adjustments. A team is more likely to collectively possess the assessment design skills for the task. It also allows for the strategic placement of Lane 1 assessments. Teachers currently do not experience sufficient support for adapting their courses.	Lesson 6: It is important to look at validity from the program perspective in order to counter compounding bias. Lesson 7: Redesigning assessment should not be a solitary task but something done collectively, to increase support and validity. The article offers an overview of required skills.
Parson (2025)	Academic misconduct is often committed because students lack an understanding of academic integrity. Offering a mandatory module for students who committed academic misconduct and a preventive course to inform and educate students on academic integrity coincided with less cases of academic misconduct.	Lesson 2: Providing preventive and punitive courses or e-learning about academic integrity can reduce cases of academic misconduct.
Perkins (2023)	"It is not the student's use of any AI tools that defines whether plagiarism or a breach of academic integrity has occurred, but whether its use is made clear by the student" (p. 7). AI can be seen as a tool for cognitive offloading.	Lesson 2: Many policies depend on the willingness of students to transparently report AI use. Lesson 4: AI offers options for cognitive offloading. The line or border of acceptable offloading will be hard to define.

Study	Findings	Interpretation and application
Perkins et al. (2024a)	The AIAS model can help teachers establish what AI usage should be authorized based on the learning goals and help to clearly communicate AI's use in each course.	Lesson 2: Clarity about what is and is not allowed can help students and teachers if combined with the redesign of assessment. Lesson 3: Several lanes can be beneficial for differentiating in how AI use is authorized per condition.
Perkins et al. (2024b)	Detection software can spot the use of AI, but with several strong caveats that might make it hard to rely on. It is advisable to organize more formative in-class assessments while at the same time redesigning assessments in which students are allowed to and informed how to use it responsibly. If institutions want better detection, they should also teach their teachers to do so.	Lesson 3: If we can't detect AI reliably, we should go for a multi-lane approach in which we have controlled exams in which AI usage is allowed. Lesson 7: Teachers could also use AI detection PD.
Sporrong et al. (2025)	AI influences current educational values. Sometimes they do not fit and undermine pedagogical or professional values, in other cases, teachers expect that AI can offer opportunities for enacting their values.	Lesson 7: Be mindful of how AI will influence important pedagogical and professional values.
Tsao (2025)	Clearer messaging, the need for adaptive policies and guidelines with ongoing student and teacher participation, availability of digital libraries of toolkits, case studies and other resources, building in early "failure experiences", and exposing students to authentic real-world applications and encounters to cultivate awareness of the limitations of GenAI.	Lesson 2: It is also important to teach students about the limitations of AI, using real-world applications and failure experiences.

Study	Findings	Interpretation and application
Walton et al. (2025)	The authors distinguish 6 types of judgments with/about AI: 1) making judgements about knowledge when working with GenAI; 2) learning to judge GenAI through its limitations; 3) relying on GenAI for things they could not otherwise do; 4) adopting ideas with low levels of criticality; 5) misjudging GenAI contributions as their own; and 6) submitting GenAI content in an assignment without judging it.	Lesson 5: There are different judgment events using AI. Students treat content and process knowledge differently, and AI does not tend to correct this artificial separation. Experience can make students more critical of AI's output, but with lower levels of disciplinary knowledge, they are less likely to find problems. Students can misattribute the work of AI to themselves. Students who use it without judging it justify this by citing its use by others and time pressure.
Waltzer et al. (2024)	Teachers were better at spotting an AI- generated essay (70%) than students (63%) or ChatGPT (60%), but all were struggling. Given that it is hard to detect, it might be smarter to redesign assessments so that they help students learn to use AI as a problem-solving tool.	Lesson 3: Either create a controlled setting to ban AI from assessment or redesign the assessment to be more demanding in other regards.
Xia et al. (2024)	"(i) teacher professional development activities for assessment, AI, and digital literacy should be provided, (ii) teachers' beliefs about human and AI assessment should be strengthened, and (iii) teachers should be innovative and holistic in their teaching to reflect the assessment transformation."	Lessons 3 and 7: In order to be able to adapt assessments to AI, teachers will require AI and assessment literacy in order to make assessments valid, reliable, and relevant, refocus them on critical thinking, and problem solving.
Zhan & Yan (2025)	Using feedback from GenAI responsibly requires evaluative judgment. A sizable part of the students engages with GenAI feedback superficially.	Lesson 5: Students need evaluative judgment in order to use GenAI responsibly. This means that they should compare and evaluate different sources of feedback including GenAI feedback and human feedback, assess their work against established criteria, and discern the change of GenAI feedback over time.