Education Research and Perspectives
An international journal

Volume 51, 2024, Pages 1-36

Generative Artificial Intelligence and Assessment Task Design: Getting Back to Basics through the Lens of the AARDVARC Model

Elaine Chapman[1†]

Jian Zhao[1]

Peyman G. P. Sabet[1,2]

1. Graduate School of Education, The University of Western Australia

2. Global Curtin, Curtin University

Effective assessments guide student learning, refine teaching practices, ensure curriculum alignment, and foster workforce readiness. However, the emergence of generative artificial intelligence (GenAI) tools, such as ChatGPT, has significantly disrupted traditional assessment processes, raising concerns about academic integrity and necessitating innovative approaches. While higher education institutions are making strides in adapting to this new reality, the foundation of effective assessment remains educators' assessment literacy. This paper responds to the critical need for improving educators' assessment literacy by introducing a comprehensive model - the 'AARDVARC' framework - that outlines eight key attributes of effective assessment: alignment, authenticity, reliability, developmental appropriateness, validity, accessibility, realism, and constructiveness. By fostering assessment literacy, educators can design innovative, equitable, and discipline-relevant assessments that incorporate GenAI responsibly and meaningfully. The paper further offers actionable recommendations for adapting university assessments to align with institutional goals and meet the evolving demands of the educational landscape. These strategies aim to ensure that assessments continue to promote student engagement, maintain academic standards, and reflect the realities of modern education.

†Address for correspondence: Graduate School of Education, The University of Western Australia, 35 Stirling Highway, Crawley, 6009, Australia. Email: elaine.chapman@uwa.edu.au

Introduction

Accurate assessments of what students know, understand, appreciate, and can do are critical in higher education. Information gathered through assessments supports various stakeholders, including students, educators, universities, policy makers and employers, in multiple ways.

The importance of assessment to various stakeholders

Assessment plays a vital role in guiding both students' learning and educators' teaching (Gibbs, 2010). For students, formative assessments are particularly valuable as they enable them to evaluate their progress, identify strengths, and pinpoint areas for improvement (Poorvu Center for Teaching and Learning, 2017). Authentic assessment tasks, such as portfolios and project-based assignments, closely mirror real-world applications, equipping students with essential professional skills (Timperley, 2020). Furthermore, self-assessment empowers students to take ownership of their learning, fostering metacognitive awareness and enhancing academic self-efficacy (Siegesmund, 2017).

For educators, assessment results provide crucial insights into students' learning and development. They help identify gaps in students' understanding so that teachers can refine their teaching strategies to address specific needs (Timperley, 2020). By analysing this data, educators can personalise instructional approaches, deliver targeted support, and enhance overall curriculum effectiveness. Moreover, assessments serve as a tool to ensure alignment between teaching objectives and learning outcomes, reinforcing the achievement of desired educational standards. Beyond individual classrooms, universities use assessment data to evaluate and refine curricula, ensuring they meet graduate attributes and accreditation requirements (Treleaven & Voola, 2008).

On a broader scale, large-scale assessments, such as standardised tests, generate data that policymakers rely on to shape education policies and allocate resources effectively, addressing disparities and promoting

equity (Tobin et al., 2015). Such assessments also provide a mechanism for institutional accountability, ensuring that educational standards are maintained and public expectations are met (Ewell, 2009).

For employers, assessment outcomes offer valuable insights into whether graduates possess critical employability skills, such as communication, teamwork, and problem-solving. When thoughtfully designed, assessments can ensure that graduates are well-prepared to meet workforce demands (Zainudden et al., 2022).

Considering the importance of assessments to all these stakeholders, the validity of assessment processes, therefore, is of paramount concern to universities. To uphold this standard, university educators must possess high levels of assessment literacy. This ensures that the tests, assignments, and other assessment tasks they design, administer, and evaluate provide accurate, reliable, and meaningful information.

Assessment against the GAI backdrop

Recent years – and in particular, since the release of ChatGPT in November 2022 (Open AI, 2022) – have seen a marked sense of panic over how university assessment systems can be adapted to cope with the ready availability of GenAI tools, such as ChatGPT (Farazouli et al., 2024; Thompson, 2023). On a worldwide basis, students, educators, academics and university leaders have grappled with question of how assessment can be done effectively against the GenAI backdrop. In particular, the appearance of potentially transformative tools such as GPT-3, which utilises an innovative large language model (LLM), as well as the new conversational interface of ChatGPT, has produced both a flurry of excitement, and further panic, amongst educators on a worldwide basis (Gamage et al., 2023).

The latest GPT-4 builds upon the foundation laid by GPT-3, introducing even more advanced capabilities in understanding and generating human-like text. GPT-4 demonstrates a significant leap in contextual comprehension, multilingual proficiency, and its ability to engage in complex, nuanced conversations (Jandhyala, 2024). These advancements have further intensified discussions within the educational

community, particularly assessment, as the potential applications and implications of such tools continue to unfold.

GenAI refers to technologies which draw upon deep learning models, generating products which resemble human responses against complex and diverse input requests (Weng, 2023). Such technologies can create systems which can address similar kinds of tasks to those traditionally addressed only through human cognition (Siemens et al., 2022). There is no question that the emergence of GenAI marks something of an inflection point with respect to university assessment processes. This is particularly true given that such tools can, with appropriate prompts, produce texts that are difficult to discriminate from human responses, even by expert assessors, as well as the very rapid and broad-scale uptake of GenAI by students, with research and social media indicating that between 33% and 50% of university students in the United Kingdom and the United States making regular use of tools such as ChatGPT to do their assessments (Adams, 2024; Nietzel, 2023).

Various higher education institutions across the world have scurried to establish new guidelines and frameworks to address the use of GenAI in assessments (Moorhouse et al., 2023; UNSW, n.d.). These frameworks primarily focus on mitigating concerns about the impact of GenAI on academic integrity, including issues of cheating and plagiarism. In a review of such frameworks at the end of 2023, Moorhouse et al. found that approximately half of the world's 50 top-ranking higher education institutions (HEI) had developed publicly available guidelines, covering three main areas: academic integrity, advice on assessment design and communicating with students. Suggested practices in such guidelines included running assessment tasks through GenAI to check the extent to which the tool can accomplish the task and having students use GenAI as part of the assessment process.

However, it is clear that university educators require more guidance to effectively integrate GenAI into their teaching and assessment practices (Lee et al., 2024; Sanusi et al., 2024). For example, researchers from the University of Adelaide revealed that while nearly half of the surveyed educators were already using AI, primarily in teaching and assessment design, less than a quarter felt adequately equipped by their institution to

use AI effectively (D. Lee et al., 2024). This underscores a significant gap in institutional support and professional development. Further research reinforces the critical need for equipping educators with not only technical knowledge, but also pedagogical strategies tailored to AI integration. Teachers have emphasised the importance of comprehensive professional development programs to ensure educators are well-prepared to adopt AI effectively in their teaching practices (Sanusi et al., 2024). These findings underscore the necessity of targeted training initiatives and institutional support to prepare educators to integrate AI in education successfully and responsibly.

Educators' Assessment Literacy

A critical component of this preparation is fostering educators' assessment literacy. Assessment literacy involves the knowledge and skills required to construct, implement, interpret, and use assessments effectively to support student learning and measure educational outcomes accurately (Pastore, 2023). Educators equipped with strong assessment literacy are better positioned to design assessments that reflect student understanding and achievement (Bayat & Rezaei, 2015) while remaining aligned with intended learning objectives (Estaji, 2024). Furthermore, assessment literacy enables educators to critically evaluate the integration of GenAI in assessments, ensuring that its use enhances educational practices rather than compromising academic integrity.

Enhancing assessment literacy enables educators to develop innovative, discipline-relevant assessment strategies that incorporate GenAI successfully and responsibly. This includes ensuring that GenAI usage is transparent, aligns with learning objectives, and promotes student engagement and equity in learning outcomes. As the educational landscape continues to evolve, strengthening educators' assessment literacy remains vital to maintaining the fairness and credibility of assessments.

Aim of this study

In this paper, we aim to respond to the critical need for improving university educators' assessment literacy by first introducing a comprehensive model of competencies for assessment literacy. This model is grounded upon eight attributes that assessment processes must

embody to ensure their effectiveness. In particular, the 'AARDVARC' model of assessment literacy is predicated on the notion that effective assessment tasks must meet all of the criteria as summarised in Table 1.

Table 1. Summary of the AARDVARC model of assessment

Attribute	**Description**
Aligned (with learning outcomes)	Prompting judgements that closely align with the intended learning outcomes for a given learning experience.
Authentic	Reflecting 'real world' challenges, which provides learners with opportunities to engage in meaningful and practical applications of their knowledge.
Reliability-promoting	Leading to consistent and dependable judgments about what students know, understand, and can do.
Developmental	Being appropriate for the specific stage in the teaching and learning cycle in which they are conducted, supporting students' ongoing growth.
Validity-promoting	Facilitating defensible inferences about what students know, understand and can do.
Accessible	Allowing all learners to demonstrate what they know, understand, and can do in an equitable way.
Realistic	Providing a practical, efficient, and sustainable means to gage what students know, understand, appreciate, and can do.
Constructive	Having positive effects on teaching and learning processes, and providing useful information for continuous quality improvement processes.

The paper has been structured as follows: we, first, discuss the 'AARDVARC' model in more detail, then provide some recommendations for adapting or developing university assessments accordingly against the backdrop of GenAI. These recommendations aim to enhance the alignment of assessments with institutional goals, learner

needs, and broader educational priorities, ensuring they serve as a catalyst for high-quality learning and teaching. Finally, a conclusions section summarises the main points discussed in this paper.

Discussion

This section discusses the eight attributes comprising the 'AARDVARC' framework in more detail as well as the questions that need to be considered in the context of GenAI.

Attribute 1: Aligned (with Learning Outcomes)

Effective assessment tasks must prompt judgements that align with the intended learning outcomes of a given learning experience. This alignment ensures that the assessments accurately evaluate whether students have achieved the desired knowledge, skills, attitudes, or understandings. To achieve this, educators need a clear understanding of what they want students to learn from the outset. *Intended learning outcomes*, or often referred to simply as, *learning outcomes*, are concise descriptions about the knowledge, understandings, attitudes, and/or skills that learners are intended to acquire during a defined learning process (e.g., learning over the course of a school term or designated part thereof).

Linking assessments to a clear set of learning outcomes not only enhances transparency for students but also reduces their anxiety. It enables them to adopt more effective approaches to learning by providing a clear framework for what is expected. For educators, clarity about learning outcomes is equally crucial. A lack of clarity can make designing effective assessment tasks significantly more challenging. Each assessment task should serve a distinct purpose, addressing one or more stipulated learning outcomes. The notion of *constructive alignment* originated by Biggs (1996) provides a simple yet effective way to think about how to align intended learning outcomes, assessment tasks, and instructional content and approaches.

Learning taxonomies are often used as a jump-off point for thinking about how to translate these learning outcomes into assessment tasks. Of these, the *Taxonomy of Educational Objectives*, or the '*Bloom's taxonomy*' (Armstrong, 2010; Gogus, 2012) is by far the most widely applied in assessment task design among universities around the world (see Shabatura, 2022; Thea, 2021). Bloom's taxonomy provides three hierarchies of educational or learning objectives, structured by complexity and specificity, within three broad domains or spheres of knowledge: the cognitive domain (i.e., mental or thinking skills); the affective domain (i.e., growth in the area of emotional responding); and the psychomotor domain (i.e., development of manual or physical skills) (Momen et al., 2022). The cognitive domain, in particular, is extensively used in education for designing assessments, as it orders cognitive skills hierarchically, requiring mastery of lower-order skills (e.g., remembering, understanding) before progressing to higher-order skills (e.g., analysing, creating).

Various questions, however, arise in the context of tools such as GenAI with respect to alignment. For example, given that factual information can now so readily be accessed through GenAI, should we simply abandon the knowledge or knowing domain entirely? Should assessments shift their focus to different kinds of knowledge, such as evaluating students' ability to locate, interpret, and apply information, rather than recalling specific facts? Furthermore, should we move away from separately assessing lower and higher cognitive levels and instead design tasks that integrate multiple cognitive dimensions - for example, implicitly assessing knowledge and understanding through activities that require creation or problem-solving? These questions challenge traditional assessment paradigms and encourage educators to think critically about how best to design assessments that remain relevant in an era of rapid technological advancement. By carefully considering these questions, educators can ensure their assessments align with learning outcomes against the backdrop of GenAI.

Attribute 2: Authentic

The second attribute in the AARDVARC model is that of authenticity. Joughin (1998) defined "authenticity" in higher education assessment context as "the extent to which assessment replicates the context of professional practice or "real life" (p. 371). In the AARDVARC model, we use this term to refer to the *extent to which students are responding to provocations that represent 'real* world' *challenges*, rather than to a specific 'type' of assessment. This might involve addressing a real-world issue, such as researching and drafting a report with recommendations for the most affordable and low-carbon transport options (McArthur, 2023). Alternatively, it might include creating business proposals, developing projects, compiling portfolios, producing artwork, or crafting videos, among other tangible and meaningful outputs (Fook & Sidhu, 2010). In other words, within the AARDVARC model, an authentic assessment is simply one in which the challenge that the student must address mirrors a reality outside of the education context. Given this, in a class on English literature, writing a book review or essay may well be deemed authentic, while the same kind of task used in another subject area (e.g., human movement) may not be so.

Ensuring that assessment tasks are authentic is always challenging (Ajjawi et al., 2024). Not all assessments within a course may meet this criterion, particularly those designed to test foundational knowledge or understanding, which serve as a prerequisite for engaging with more complex, application-based tasks. To design authentic assessments, educators must clearly define the desired end-outcomes relevant to their specific disciplines - a task made even more difficult in a rapidly evolving world. Educators need to be well-versed in a diverse range of assessment strategies, enabling them to select approaches that best align with particular learning outcomes.

The rapid evolution of GenAI has added another layer of complexity to this challenge. In many disciplines, educators must now prepare their students for future work and personal lives that no one can fully envisage at this point. In this context, accurately anticipating the skills and

knowledge graduates will need to thrive as productive members of society is an extraordinarily difficult undertaking.

In order to embrace GenAI in a meaningful way, therefore, educators need not only to become well-versed in the use of such tools as they stand presently, but also keep abreast of ongoing advancements to stay informed about how the tools might be used in their disciplines in future (A. V. Y. Lee, 2024). Concurrently, however, authentic assessment can also help educators to ensure that their tasks continue to produce valid judgements about what students know, understand and can do. By incorporating activities such as fieldwork, addressing pressing local issues, or tackling real-world problems in collaborative work groups, educators can design assessments that both prepare students for complex challenges and resist over-reliance on GenAI. These approaches encourage deeper engagement, critical thinking, and the application of knowledge in ways that cannot be fully replicated by AI tools alone.

Attribute 3: Reliability-Promoting

The third attribute is that of promoting reliable scores or judgements. Effective assessment tasks prompt consistent judgements about what students know, understand, and can do. In the assessment literature, the term *reliability* is used to refer to the *consistency* (alternatively, the *stability* or dependability) of scores or judgements that arise from assessment evidence (Reynolds et al., 2021). This means that a student's score on an assessment should not be influenced by irrelevant or random factors such as *when* the student undertakes the assessment (morning vs. afternoon), or *who* makes the judgement.

Therefore, in terms of assessment task design, educators need to focus upon whether a given task will allow them to make dependable judgements about what students know, understand, and can do. Almost anything can be a source of inconsistency or instability in assessment scores. For example, inconsistencies across testing occasions occur when scores depend on the *timing* students take the task. Inconsistencies across task *forms* happen when results are influenced by the version of the task

or the options chosen. Inconsistencies between raters or markers arise when scores vary based on *who* made the judgement.

The implications of GenAI for designing tasks that promote dependable or reliable judgements really come down to the nature of the tasks that are likely to be used in this context. Challenges associated with achieving dependability vary widely across these different kinds of assessment tasks. The use of what are called *selected-response* questions (in which students choose a response from a list of alternatives, such as in a multiple-choice test item) within examinations or quizzes essentially rules out differences in judgements that may arise across different markers. In *constructed-response* tasks, where the student must construct their response based on a broad prompt provided to them, it can be very difficult to avoid sources of inconsistency such as differences between markers entirely.

Achieving consistency across markers can be particularly tricky to achieve in the context of the kinds of complex performance-based assessments (one form of a constructed-response task) that are likely to be used in the GenAI context. In these tasks, different samples of behaviour (performance) may be judged by different markers, at different times. In light of this, the GenAI context will call upon educators to become highly rigorous in the application of tools such as relevant rubrics to ensure that the judgements made on the basis of assessment evidence are consistent.

Attribute 4: Developmental

Effective assessment tasks also prompt judgements that are appropriate for the timepoint in the teaching and learning cycle in which they will be completed. Traditionally, assessments in school and university education have tended to focus on the assessment *of* learning - that is, focus on assessing only what students know, understand and can do at the *end* of a course or bounded learning experience (Amua-Sekyi, 2016; Swiecki et al., 2022). More recently, however, there has been a call for greater emphasis to be placed on the use of assessments *for* and *as* learning (Dann, 2014; Rutherford et al., 2024). The term **assessment for learning**

refers to educators gathering evidence on students' knowledge, understanding and skills to inform their own teaching practices. This category can include both 'diagnostic' assessments and also 'formative' assessments. **Assessment as learning** refers primarily to assessments in which students act as their own assessors (self-assessment), reflect upon their own work (self-reflection), or assess the work of other students (peer or small group assessment).

It is important to consider the function that will be served by any particular task and to take this into account in the design of that task. For example, in diagnostic assessments with an education context, the tasks used should be particularly helpful for identifying whether the students have the prerequisite knowledge or understandings to attempt a new learning experience, and also, to identify any misconceptions that need to be addressed (Marchant, 2023). An example of a tasks that could be useful here is a concept map, which focuses on the links that students are perceiving between concepts they are learning (Jackson et al., 2024).

In assessment *as* learning tasks such as self-reflection activities, educators may choose to focus not on whether students arrive at the *right* answers in their evaluations, but more upon *how they arrived at these* and their ability to reflect upon their own learning. Reflective diaries, evaluated on a pass/fail basis, are well-suited for this purpose as they encourage deep self-assessment without the pressure of achieving a "correct" answer. For summative assessments, which involve higher-stakes decision-making (Kibble, 2017), the technical soundness of the tasks and the reliability of the associated marking processes are paramount. These assessments require rigorously designed tasks that ensure validity and consistency, reflecting their critical role in evaluating overall achievement.

The context of GenAI further underscores the importance of using formative tasks in overall assessment plans. Ongoing tasks that require students to produce work iteratively and improve based on feedback are significantly less prone to 'faking' compared to one-off tasks. Furthermore, integrating GenAI tools into assessments is essential to ensure assessments mirror real-world contexts. Formative assessments can achieve this by moving beyond asking students to replicate what these tools can already accomplish. Instead, these tasks should focus on

enabling students to leverage the outputs of such tools to meet specific objectives, ensuring that they develop their knowledge and understanding in a meaningful manner.

Attribute 5: Validity-Promoting

Effective assessment tasks prompt defensible inferences about what students know, understand and can do. The term **validity** refers to the extent to which an assessment mechanism yields appropriate inferences about a given *latent construct*. In other words, it is the "inferences regarding specific uses of a test [that] are validated, not the test itself" (American Psychological Association, 1985, p. 9). At a minimum, to achieve a high level of validity, our assessment tasks must include good *indicators* of our intended learning outcomes (Eignor, 2013), and we must also interpret students' responses to this task in a defensible way (Kladaras & Haudek, 2022). For example, if we attempted to assess students' ability to conduct and interpret the products of particular statistical procedures, did the scores we awarded provide an accurate picture of whether students could do this, or did they reflect the influence of other factors, such as students' reading ability?

There are two main threats to the validity of any given assessment task: construct *underrepresentation*, and construct *irrelevance*. The former threat arises when given assessment scores underrepresent the learning outcomes being targeted (e.g., the tasks are too 'narrow' to fully represent the intended learning outcomes). This can happen, for example, when tasks only focus upon a portion of a semester's worth of learning. The latter threat arises when factors that are irrelevant to the outcomes being targeted have a reliable and significant influence on assessment scores. For example, if an educator intends *only* to measure students' ability to read research articles in a critical way, but students are required to work in a group and provide an oral presentation to demonstrate this ability, the scores that arise from this may be confounded reliably with other factors, such as social anxiety and communication skills. If, of course, being able to work in a group and communicate effectively in oral form are important generic learning outcomes that the educator is *also* seeking to assess, the influence of these factors is not irrelevant. In simple terms,

underrepresentation occurs when the test measures less than the construct it is planned to measure, and irrelevance refers to when a test measures characteristics, skills or content that are not related to the test construct (AERA et al., 1999).

Based on the example above, the degree to which any given task promotes valid inferences about what students know, understand and can do will depend entirely on what is seen to be relevant or irrelevant within a particular context. There are currently no guidelines for addressing the validity of GenAI-based assessments and their results (Kaladaras, et al., 2024). Therefore, educators need to re-evaluate this question in depth. For example, given that GAI is likely to be used in the workplaces of many graduates in the future, to what extent should the ability to use these tools in a meaningful, discipline-relevant way be seen as a generic attribute that is relevant across all assessment tasks?

Attribute 6: Accessible

Effective assessment tasks should be inclusive and allow all learners to demonstrate what they know, understand, and can do in an equitable way. The term 'accessible assessment' has come to mean many different things depending upon the context in which the term is used, e.g., "a lack of 'hard' access capabilities" (Roelofs, 2019, p. 22), assessment design flaws (Beddow et al., 2008) or limited access to resources required for the completion of the assessment task.

Assessment accessibility is not a static property of a test but is instead the result of the interaction between the test-taker's characteristics and test features that affect the test-taker's performance in the test (Kettler et al., 2009; Winter et al., 2006). It is, therefore, considered as a prerequisite to validity (Kane, 2004; Roelofs, 2019). In our use of this term, we are referring to assessments that permit all students to demonstrate what they know, understand, appreciate, and can do, including those in specific minority groups and with particular additional support needs.

The notion of assessment accessibility is related to the notion of assessment bias, which is present whenever one or more items on a test offend or unfairly penalize students because of those students' personal characteristics, including their race, gender, socioeconomic status, or

religion. Equitable, inclusive and accessible assessment mechanisms are the antithesis of biased assessment practices. Popham (2006) noted that assessment items can invite bias into assessment judgements via their content (e.g., items that include content that is different or unfamiliar to different respondents); language (e.g., items that include group-specific language, vocabulary, or reference pronouns); and structure (e.g., items which include ambiguities that benefit or disadvantage certain respondents or groups of respondents). Another source of bias in assessment tasks relates to the *judgements* that educators make in marking the assessment responses, which can be influenced by unconscious biases that impact the judgements made, including anchoring and confirmation biases (Reynolds et al., 2009).

GenAI and other emerging tools have significant potential to introduce additional inequities in assessment processes. If assessment tasks can or will involve the use of such tools, students' success in given activities may depend heavily on factors such as the level of access they have to these tools and also their skills in making use of them. This potential underscores the importance of taking a transparent approach to the use of GenAI in assessment tasks, providing all students with access to the tools, and, where relevant, to training in their use, to 'relevel' the playing field.

Attribute 7: Realistic

Effective assessment tasks are also practical, efficient, and sustainable. Assessment tasks that cannot be sustained realistically will not produce the kind of durable positive impacts that we are seeking through the use of effective assessment mechanisms. For example, recent years have seen an increasing number of concerns about overassessment in education. Particular concerns have been raised about potential negative impacts of excessive assessment workloads on student learning outcomes (Ediger 2022; Tomas & Jessop, 2018). Cited negative effects for students have included reduced motivation levels (Kusurkar et al., 2023), increases in academic stress levels (Kenwright, 2018) and increased temptations for students to engage in academic misconduct (Draper et al., 2022). Various studies have also documented negative

effects on staff from overassessment, including reducing the time they have to plan lesson content and pedagogies (Parry et al., 2019), leading ultimately to teacher burnout.

Different assessment tasks can vary widely in terms of cost, which should be a significant factor in any choices made amongst particular assessment strategies. Costs can include financial resources (e.g., purchasing licenses for specialised software), but can also include costs in terms of educators' time (in designing and implementing the task, or in marking students' responses to it), and opportunity costs for students (in terms of other activities they could be undertaking instead). All assessment tasks will also, however, have alternatives, and it falls to educators to determine which of these options is the most practical and efficient in given circumstances.

GenAI and other such tools have significant potential to enhance the efficiency of assessment tasks. For example, there is significant potential to use AI-driven adaptive learning systems to provide automated grading of some assessment tasks (Gnanaprakasam & Lourdusamy, 2024). This can not only free up educators to focus on other elements of the teaching and learning process but can also enrich students' learning experiences by offer personalised feedback that would not otherwise be attainable. GenAI systems also have limited abilities to identify gaps in learning, which educators can then harness in a diagnostic capacity.

Attribute 8: Constructive

The final attribute in the AARDVARC model underscores the need for assessment tasks to have positive effects on teaching and learning processes, and provide information that is useful in continuous quality improvement processes. One of the manifestations of this attribute, albeit limited in scope, is the 'positive washback' effect, defined as the positive "effect of testing on teaching or learning" (Hughes, 2003, p.1).

In a sense, to be constructive, assessments necessarily need to embody all of the former attributes in the model. Being constructive, however, requires going beyond prompting reliable and valid judgements about

what students can do in the context of authentic, aligned, accessible and realistic tasks. Being constructive requires educational institutions and teachers to use the information that is collected through assessments to monitor, diagnose and problem-solve how best to foster individual students' learning. Being constructive, therefore, goes beyond the previous attributes, and calls upon educators to use different forms of data to create powerful learning environments.

Again, GenAI has significant potential to enhance the degree to which assessment data contributes constructively to student learning processes. For example, GenAI can be used to provide immediate or near-immediate summaries and analyses of students' responses to short assessment items within classes. This means that students can receive very rapid feedback in developing their understandings. Such information can also provide educators with timely information on students' understandings of given concepts, allowing them to re-teach critical concepts before moving on to more advanced topics. As a result, education contexts can become both more interactive and more effective.

Recommendations for the use of GenAI in assessment

In this section, we provide a few suggestions for educators to address the challenges introduced by GenAI in maintaining the academic integrity of educational assessment practices aligning with the AARDVARC attributes. It is important to note that this is just a starting point, and numerous other strategies may also prove effective in this context. To achieve this, educators can:

1. Incorporate real-life examples, contextually specific situations and current and evolving topics into assessment tasks (Authentic)
Designing assessment tasks around real-life examples or unique, context-specific examples not only reduces the likelihood of AI-generated outputs but also compels students to engage deeply with the material and produce tailored responses (The University of Melbourne, 2023b). For example, incorporating case studies that include real-life examples or local contexts encourages students to synthesise knowledge creatively. These tasks are challenging for AI tools, as their outputs are limited by

pre-trained data and algorithmic patterns, which struggle to address highly specific or nuanced scenarios (Rana et al., 2024), limiting their ability to address highly specific or novel scenarios.

Assessment tasks that focus on analysing very recent events or developments not yet extensively represented in GenAI training data remain an effective strategy. GenAI models are trained on data up to a specific cut-off date, and while recent reports suggest that tools like ChatGPT may access up-to-date information, no definitive evidence confirms this capability (Radford & Kleinman, 2023). At present, this approach encourages students to engage with current events and critically examine contemporary issues, thereby reducing reliance on pre-trained AI responses and fostering independent thought.

2. Use authentic assessment tasks (Authentic)

Authentic assessment tasks offer students opportunities to engage with real-world problems and scenarios (Ajjawi et al., 2020a; Kaider et al., 2017), bridging the gap between academic learning and professional practice. These tasks, such as fieldwork project analysis, case study, group work and simulations and virtual labs (Bosco & Ferns, 2014; Nguyen, 2023), require students to apply their knowledge in practical contexts, fostering skills that are directly transferable to their future careers.

By closely mirroring the demands of professional environments, authentic assessments enhance the relevance and applicability of learning while making it more difficult to rely on AI-generated responses, as they usually require "critical thinking, collaboration, and ethical reasoning" (Awadallah Alkouk & Khlaif, 2024, p. 3). Moreover, authentic assessments encourage students to actively solve real-world problems, apply their knowledge, and make informed decisions, fostering the development of both cognitive and metacognitive skills (Ajjawi et al., 2020b).

While sophisticated designs of authentic assessments that address the complexities of the digital world remain underexplored (Bearman & Ajjawi, 2023), the integration of such elements is poised to become a defining trend in the development of future assessment tasks.

3. Incorporate the use of self- and peer assessments into overall assessment plans (Developmental)

Self- and peer assessment actively engages students in evaluating their own work and that of their peers, promoting deeper learning and reflection. According to Thomas et al. (2011), self-assessment enhances students' learning outcomes by requiring them to make sophisticated judgments about their performance and understanding. By engaging in self-assessment, students develop a realistic perspective of their strengths and areas for improvement, fostering self-regulation and accountability. Peer assessment, on the other hand, encourages collaborative learning and critical thinking (Karandinou, 2012; Kollar & Fischer, 2010). By evaluating their peers' work against established criteria, students gain valuable insights that can inform and enhance their own work.

The primary aim of peer and self-assessment is not merely to assign grades to one's own work or that of peers but to deepen understanding of learning objectives and quality standards, develop reflective and evaluative skills, and promote self-regulated learning (Stancic, 2020). This approach transforms assignments from being solely outcome-focused to evaluating the processes of learning and improvement (Wanner & Palmer, 2018). This shift encourages active engagement and reduces the likelihood of students relying solely on GenAI tools, as such assessments require personalised input and critical analysis. For instance, self-assessment requires students to critically evaluate their efforts, aligning their work with learning objectives and course criteria. Peer assessment further enriches the learning process by exposing students to diverse perspectives and approaches, enhancing their ability to critique and refine ideas effectively.

4. Require students to refer to lectures/tutorials or in-class discussions (Constructive)

Incorporating requirements for students to reference specific lectures, tutorials, or in-class discussions in their assessments, such as essays or reflective writing tasks, can effectively reduce reliance on GenAI. Since GenAI models are not typically trained on such course-specific content,

requiring these references ensures that students engage with materials unique to their educational context. This strategy compels students to actively participate in and draw from their learning experiences, making their work more personalised and less susceptible to being outsourced to AI tools.

By embedding this requirement, educators challenge students to connect theoretical knowledge discussed in class with their own interpretations and arguments. For example, an essay could prompt students to analyse a case study discussed in a tutorial, integrating key points raised during the session. This not only fosters deeper engagement with the material but also encourages students to apply what they have learned in meaningful ways, strengthening their critical thinking and analytical skills.

Moreover, this strategy enhances classroom engagement, as students recognise the direct relevance of discussions and lectures to their assessments. It motivates active participation and attentiveness, fostering a stronger connection between their learning process and the outcomes they are evaluated on. Requiring references to course-specific content also provides educators with clearer insights into students' comprehension and ability to synthesise information, making assessments more meaningful and reflective of genuine learning.

5. Integrate interactive and iterative assessments (Developmental)
Interactive and iterative assessments are essential for fostering deep learning and reducing students' reliance on GenAI. These approaches create opportunities for students to actively engage with the learning process and demonstrate their understanding in dynamic, iterative ways that are difficult for GenAI to replicate.

Incorporating real-time, interactive assessments such as in-class presentations, debates, or group discussions requires students to articulate their reasoning and *justify* conclusions. These tasks demand spontaneous responses and active engagement, making it more difficult for students to complete assessments using GenAI (The University of Melbourne, 2023c). For example, students presenting their problem-solving approaches must answer live questions, explain their thought

processes, and adapt their responses in situ. This encourages critical thinking, clarity of communication, and a deeper grasp of the subject matter.

Iterative assessments or staged assessments involve multiple stages where students refine their work based on feedback, making it more challenging for students to complete the task using GenAI by incorporating elements such as group work and requiring reflections on each student's specific individual contributions (The University of Melbourne, 2023a). Tasks like research projects, portfolios, or case study analyses can be structured into phases - proposal, draft, and final submission - with feedback guiding the evolution of the work. This iterative approach not only strengthens students' understanding of the material but also develops their ability to critically evaluate and improve their outputs over time. By engaging in this process, students build skills in self-regulation and reflection, reducing the temptation to rely on quick fixes from AI tools.

By leveraging interactive and iterative assessments, educators can design tasks that promote deeper engagement, critical thinking, and a commitment to continuous improvement, all while maintaining the integrity of the assessment process.

6. *Have students critique GenAI outputs (Developmental and Accessible)*

Encouraging students to critically evaluate GenAI outputs offers a valuable opportunity to develop analytical skills and deepen their understanding of both the subject matter and the capabilities of AI tools.

A university in Singapore has already integrated the use of ChatGPT into this workshop to teach students how to effectively utilise and critically evaluate AI outputs (C. Lee & Low, 2024). Tasks that require students to assess the quality, accuracy, and coherence of AI-generated responses help foster critical thinking and promote a nuanced awareness of the strengths and limitations of GenAI. For example, students might be tasked with comparing an AI-generated essay or solution to academic standards, evaluating its effectiveness in conveying ideas, accuracy in addressing the prompt, and adherence to logical reasoning.

This approach helps students become informed users of GenAI, teaching them to recognise when these tools may be useful and where they fall short. By engaging in this critique process, students learn to identify issues such as biases, oversimplifications, or a lack of depth in AI-generated content. Moreover, critiquing AI outputs also aligns with authentic learning practices, as students must apply their knowledge to assess and improve existing content. For instance, they could be asked to revise an AI-generated response to make it more accurate, relevant, or contextually appropriate, further demonstrating their understanding of the material. This approach not only enhances learning outcomes but also equips students with skills to critically navigate the increasing presence of GenAI in academic and professional settings.

Whilst GenAI has introduced a variety of new challenges to educators in terms of maintaining academic integrity and accessibility, it has also introduced a host of new possibilities for enhancing assessment efficiency and efficacy. Some of the ways in which educators can use GenAI and other emerging tools to enhance assessment processes aligning with the AARVARC attributes include:

1. Using GenAI to assist in the design of novel assessment materials or input prompts (Aligned, Authentic and Developmental)
GenAI tools can be used to develop prompts for creative writing with varying difficulty levels or to design case study scenarios with varying degrees of complexity. The capabilities of these tools to simulate authentic work environments in the scenarios should be optimally exploited. Such scenarios can be effective resources in which to develop students' workplace skills such as critical thinking and problem-solving.

An example of using GenAI to design novel assessments is a 'Case Study Analysis' in which GenAI generates several case studies with varying complexity levels. In a simple case study, learners will be asked to analyse an institution's financial statements, while in a complex case study, they will be involved in developing a strategic plan for the launch of a new service or product.

To receive optimum outputs, users need to feed into GenAI tools precise statements or questions, known as prompts. To generate effective prompts, Harvard University (2024) advises prompt writers to:

I. Be specific
II. Act as if the AI tool were a person or an object
III. Advise AI how they would like the output to be presented.
IV. Use 'do' and 'don't'.
V. Give AI examples that resemble the expected output.
VI. Give AI some details about the audience and the sort of tone expected in the output.
VII. Develop prompts in a reiterative process. Start with basic questions and add to them over time. Revise the tone or the wording or add more details gradually.
VIII. Interact with the AI tool as if you are working together on a project and give feedback. Tell it which parts of the output need to be improved.
IX. Ask it to help you by creating a prompt for you. For example, "How should I ask you for help with writing an argumentative essay on the use of AI in education?" and gradually add to it at each stage.

2. Providing automated marking of students' responses (e.g., on criteria such as grammatical and syntax errors), as well as personalised, rapid feedback (Realistic, Reliable, Aligned and Constructive)

GenAI is able to quickly and accurately mark selected response items, reducing teachers' marking load and leaving them with more time to spend on other academic commitments such as lesson planning.

GenAI can also be used to generate questions that are tailored to learners' learning needs or preferences. For example, if a learner is struggling with the use of prepositions in English, GenAI can create a grammar lesson focusing on the use of prepositions, while a student who can use prepositions accurately but has problems with subject-verb agreement will be given a task focusing on this grammatical area.

GenAI can provide students with personalised feedback following the analysis of their written works. The feedback will address their errors and highlight the gap in their knowledge. For instance, if the learner consistently makes errors in the use of collocations in English, GenAI can suggest activities that facilitate the accurate use of collocations.

3. Developing starting rubrics (Aligned and Reliable)
GenAI can also help with developing a starting rubric in several ways. The first possibility is to arrive at a basic rubric structure by feeding in prompts comprised of detailed assessment information such as learning objectives and content. GenAI can also tailor the already existing rubrics to learners' specific needs such as the difficulty level.

The other option for using GenAI in the development of starting rubrics is to have an existing rubric analysed by the GenAI tool and the inherent gaps or inconsistencies identified. GenAI can also help to make descriptors measurable by suggesting appropriate action verbs and providing clear examples.

4. Developing starting multiple-choice questions and short-form (short answer) quiz questions based on specific inputs (Aligned, Reliable and Realistic)
GenAI can create a diverse range of question types such as short answer and multiple-choice questions based on given learning objectives, text or a theme. In the same way, GenAI can adjust the difficulty level of questions, tailor assessments to learners' learning preferences or make them more personalised. It is also possible to receive help from GenAI in developing plausible and engaging distractors for multiple-choice question types. In addition, GenAI can present a sequence of questions in a progressively difficult order. With the help of GenAI, it will be possible to create question banks that allow for easy access to questions. GenAI can also facilitate decisions about the number of questions required to accurately assess a learner's knowledge.

5. Generating starter discussion prompts (Valid and Authentic)
GenAI can contribute to creating engaging discussion questions by using a diverse range of prompts. These prompts can be created by feeding in GenAI a topic such as "global warming", resulting in various discussion

questions ranging from factual-based to opinion-based sources. To add more diversity to these types of questions, GenAI can include prompts that consider multiple perspectives and enhance critical thinking.

GenAI can also provide text-based prompts by receiving a text and creating discussion questions centring around the content from an analytical perspective, the author's viewpoints and the possible implications. In the same way, GenAI can compare and contrast different texts and create prompts that can develop analytical skills and promote critical thinking. GenAI can also develop scenario-based prompts through creating hypothetical scenarios that can provoke discussion and enhance critical thinking.

6. Generating starter case studies for use as assessment prompts (Aligned, Authentic and Constructive)

Thanks to its generative power, GenAI can create a multitude of scenarios across different disciplines that are based on diverse realistic situations and with varying levels of complexity. GenAI has the capability of integrating multiple real-world data sets to create authentic case studies. For example, to expose students to an authentic situation, GenAI can integrate the current industry trend with the challenges it is facing. GenAI can also create engaging personalised scenarios by analysing learners' interests and identifying their learning styles and preferences.

GenAI can develop scenarios that promote critical thinking and creativity in different ways. For example, it can create open-ended scenarios that foster learners' creativity and enhance their problem-solving skill, or it can develop ethical dilemmas that require learners to make calculated decisions after considering multiple perspectives.

7. Generating alternative examination or assignment questions based on previous prompts used (Aligned and Constructive)

GenAI can be used to analyse a learner's strengths and weaknesses, and their preferred learning style to create tailored assessments. GenAI tools can be used to generate questions that align with the learner's specific needs. GenAI also allows for adjusting the difficulty levels of assessment based on a learner's performance. This can be used to ensure that

assessments are always challenging but not overwhelming, a factor that can promote learning.

8. Assisting to explain why certain students' responses are incorrect or less than satisfactory (Developmental and Realistic)
GenAI can compare leaner's answers with correct responses determined by experts and identify deviations in learner responses. GenAI can also explain why each correct answer is correct, developing students' understanding of the underlying reasons. GenAI can analyse a large corpus of learner responses to a specific question and an assignment and identify the common mistakes, misconceptions and difficult areas.

Conclusions

In light of the current rapid transformation that education is undergoing, the present paper aimed to address the accompanying need for changes in assessment task design through the lens of the AARDVAC model. New technologies will have important implications for all of the attributes in the AARDVARC model. While many of the shifts that such tools have introduced can be framed as challenges for educators, these tools also have significant potential to enhance the accuracy, efficiency and efficacy of assessment processes at all levels of education. Using models such as the AARVARC can provide a systematic basis for thinking about how to address these challenges and ensure that we continue to design effective assessments that yield valid judgements, irrespective of the evolving contexts in which we work.

Authors

A/Professor Elaine Chapman holds a PhD in Psychology and has over 20 years of teaching experience across three leading universities - Monash University, the University of Sydney, and the University of Western Australia. Her teaching expertise spans child and educational psychology, assessment, quantitative research design, and statistics. She serves as an editor for high-impact journals, including *Frontiers in Psychology* and the *British Journal of Educational Psychology*. She has delivered invited video presentations for SAGE on quantitative methods in education and

has provided keynote addresses at conferences centred on measurement and assessment.

Dr. Jian Zhao is an early-career researcher at the Graduate School of Education, the University of Western Australia. Specialising in mental health measurement, assessment design, and mixed-methods research, she has contributed to projects on the mental health of Chinese international students in Australia and the prediction of self-harm, suicidal behaviours among young people in Western Australia. Her key research interests include mental health measurement, assessment design, international student mental health, coping strategies, resilience, and the prediction of suicide and self-harm.

Dr. Peyman G.P. Sabet is a Doctor of Education candidate at the University of Western Australia with a focus on educational psychology and the internationalisation of Australian tertiary education. He also holds a PhD in language and intercultural education from Curtin University where he works as a lecturer in TESOL. Peyman has been involved in language pedagogy and linguistics for more than twenty-five years, with a wealth of teaching experience and publications in a number of peer-reviewed journals. Peyman's research expertise lies in the areas of Interlanguage Pragmatics, Inter/Cross-cultural Communication, Intercultural Competence, Assessment, Vague Language and second language acquisition.

References

Adams, R. (2024, February 1). More than half of UK undergraduates say they use AI to help with essays. *The Guardian*. https://www.theguardian.com/technology/2024/feb/01/more-than-half-uk-undergraduates-ai-essays-artificial-intelligence

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment and Evaluation in Higher Education*, *49*(4), 499–510. Scopus. https://doi.org/10.1080/02602938.2023.2271193

Ajjawi, R., Tai, J., Huu Nghia, T. L., Boud, D., Johnson, L., & Patrick, C.-J. (2020a). Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment & Evaluation in Higher Education*, *45*(2), 304–316. https://doi.org/10.1080/02602938.2019.1639613

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* American Psychological Association.

Amua-Sekyi, D. E. T. (2016). Assessment, student learning and classroom practice: A review. *Journal of Education and Practice, 7*(21), 1–6. https://files.eric.ed.gov/fulltext/EJ1109385.pdf

Armstrong, P. (2010). *Bloom's Taxonomy*. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Awadallah Alkouk, W., & Khlaif, Z. N. (2024). AI-resistant assessments in higher education: Practical insights from faculty training workshops. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1499495

Bayat, K., & Rezaei, A. (2015). Importance of teachers' assessment literacy. *International Journal of English Language Education*, *3*, 139. https://doi.org/10.5296/ijele.v3i1.6887

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test accessibility and modification inventory*. Vanderbilt University.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://link.springer.com/article/10.1007/BF00138871

Bosco, A. M., & Ferns, S. (2014). Embedding of authentic assessment in work-integrated learning curriculum. *Asia-Pacific Journal of Cooperative Education*, *15*(4), 281–290. https://www.ijwil.org/files/APJCE_15_4_281_290.pdf

Dann, R. (2014). Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 149–166. https://doi.org/10.1080/0969594X.2014.898128

Draper, M., Perry, A., & Berry, J. (2022, May 4). *Academic integrity, blended learning and virtual delivery–sector responses in the context of an institutional case study*[Paper presentation]. 8th European Conference on Academic Integrity and Plagiarism: Ethics and Integrity in the Changing World. Porto. Portugal. efaidnbmnnnibpcajpcglclefindmkaj/https://academicintegrity.eu/conference/wpcontent/files/2022/Book_of_Abstracts_2022.pdf.

Ediger, M. (2022). Excessive testing and pupils in the public schools. *A Journal Pertaining to College Students*, 168. https://eric.ed.gov/?id=EJ1143723

Eignor, D. R. (2013). The standards for educational and psychological testing. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, & S. P. Reise, et al. (Eds), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment: Industrial and organizational psychology*. American Psychological Association.

Estaji, M. (2024). Perceived need for a teacher education course on assessment literacy development: Insights from EAP instructors. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*(1), 50. https://doi.org/10.1186/s40862-024-00272-2

Ewell, P. (2009). *Assessment, accountability and improvement: revisting the tension*. National Institute for Learning Outcomes Assessment.

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, *49*(3), 363–375. https://doi.org/10.1080/02602938.2023.2241676

Fook, C. Y., & Sidhu, G. K. (2010). Authentic assessment and pedagogical strategies in higher education. *Journal of Social Sciences*, *6*(2), 153–161.

Poorvu Center for Teaching and Learning *(2017). Formative and Summative Assessments*. https://poorvucenter.yale.edu/Formative-Summative-Assessments

Gamage, K. A. A., Dehideniya, S. C. P., Xu, Z., & Tang, X. (2023). ChatGPT and higher education assessments: More opportunities than concerns? *Journal of Applied Learning and Teaching*, *6*(2), 358–369. Scopus. https://doi.org/10.37074/jalt.2023.6.2.32

Gibbs, G. (2010). *Using assessment to support student learning*. Leeds Met Press. https://eprints.leedsbeckett.ac.uk/id/eprint/2835/

Gogus, A. (2012). Bloom's taxonomy of learning objectives. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 469–473). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_141

Harvard University Press. (2024). https://huit.harvard.edu/news/ai-prompts

Huges, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Jackson, I., Ivanov, D., Dolgui, A., & Namdar, J. (2024). Generative artificial intelligence in supply chain and operations management: A capability-based framework for analysis and implementation. *International Journal of Production Research*, *62*(17), 6120–6145. Scopus. https://doi.org/10.1080/00207543.2024.2309309

Joughin, G. (1998). Dimensions of oral assessment. *Assessment and Evaluation in Higher Education*, *23*(4), 367–378. https://web.mit.edu/jrankin/www/oral_exams/joughin.pdf

Kaider, F., Hains-Wesson, R., & Young, K. (2017). Practical typology of authentic work-integrated learning activities and assessments. *Asia-Pacific Journal of Cooperative Education*, *18*(2), 153–165. https://eric.ed.gov/?id=EJ1151141

Kaladaras, L., Akaeze, H. O., &. Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence.

Frontiers in Education 9. https://doi.org/10.3389/feduc.2024.1399377

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next-generation science standards performance assessments. *Frontiers of Education 7*. https://doi.org/10.3389/feduc.2022.96828

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspectives, 2,* 135–170. https://doi.org/10.1207/s15366359mea0203_1

Karandinou, A. (2012). Peer-assessment as a process for enhancing critical thinking and learning in design disciplines. *Transactions*, *9*, 53–67. https://doi.org/10.11120/tran.2012.09010053

Kenwright, B. (2018). Managing stress in education. *Frontiers in Education, 1-8*. https://www.xbdev.net/misc_demos/demos/managing-stress-education/paper.pdf

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*, 529–551. https://doi.org/10.1080/01619560903240996

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *41*(1), 110–119. https://doi.org/10.1152/advan.00116.2016

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*, 344–348. https://doi.org/10.1016/j.learninstruc.2009.08.005

Kusurkar, R. A., Orsini, C., Somra, S., Artino, A. R., Daelmans, H. E.M., Schoonmade, L. J., & van der Vleuten. (2023). The Effect of assessments on student motivation for learning and its outcomes in health professions education: A review and realist synthesis. *Academic Medicine 98*(9), 1083-1092. https://doi.org/10.1097/ACM.0000000000005263

Lee, A. V. Y. (2024). Staying ahead with generative artificial intelligence for learning: Navigating challenges and opportunities with 5Ts and 3Rs. *Asia Pacific Journal of*

Education, *44*(1), 81–93. Scopus. https://doi.org/10.1080/02188791.2024.2305171

Lee, C., & Low, M. Y. H. (2024). Using genAI in education: The case for critical thinking. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1452131

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence, 6, 100221*. https://doi.org/10.1016/j.caeai.2024.100221

Marchant, R. (2023, April 3). Diagnostic assessments—Teacher's guide & examples. ICAS Assessments. https://www.icasassessments.com/blog/diagnostic-assessments-teachers-guide-examples/

McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. *Higher Education, 85*(1), 85–101. https://doi.org/10.1007/s10734-022-00822-y

Momen, A., Ebrahimi, M., & Hassan, A. (2022). Importance and implications of theory of bloom's taxonomy in different fields of education (pp. 515–525). https://doi.org/10.1007/978-3-031-20429-6_47

Momen, A., Ebrahimi, M., Hassan, A.M. (2023). Importance and implications of theory of bloom's taxonomy in different fields of education. In: Al-Sharafi, M.A., Al-Emran, M., Al-Kabi, M.N., Shaalan, K. (eds) *Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems*. *ICETIS 2022. Lecture Notes in Networks and Systems, vol 573*. Springer, Cham. https://doi.org/10.1007/978-3-031-20429-6_47

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

Nguyen, N. (2023). *What is authentic assessment? A full guide*. Feedback Fruits. https://feedbackfruits.com/blog/what-is-authentic-assessment-a-full-guide-for-educators

Nietzel, M. T. (2023, March 20). *More Than Half Of College Students Believe Using ChatGPT To Complete Assignments Is Cheating*. Forbes. https://www.forbes.com/sites/michaeltnietzel/2023/03/20/more-than-half-of-college-students-believe-using-chatgpt-to-complete-assignments-is-cheating/

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

Parry, S., Allen, M., & Briten, E. (2019). Where has all the science gone? *Primary Science, 157*, 14-15. https://eric.ed.gov/?id=EJ1252143

Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1217167

Radford, A., & Kleinman, Z. (2023, September 27). *ChatGPT can now access up to date information*. https://www.bbc.com/news/technology-66940771

Rana, N. P., Pillai, R., Sivathanu, B., & Malik, N. (2024). Assessing the nexus of Generative AI adoption, ethical considerations and organizational performance. *Technovation*, *135*, 103064. https://doi.org/10.1016/j.technovation.2024.103064

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). Reliability. In C. R. Reynolds, R. A. Altmann, & D. N. Allen (Eds.), *Mastering Modern Psychological Testing: Theory and Methods* (pp. 133–184). Springer International Publishing. https://doi.org/10.1007/978-3-030-59455-8_4

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment* (2nd ed.). Pearson.

Roelofs, E. (2019). A framework for improving the accessibility of assessment tasks. In B. P. Veldkamp and C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 21-45). Springer.

Rutherford, S., Pritchard, C., & Francis, N. (2024). Assessment IS learning: Developing a student-centred approach for assessment in Higher Education. *FEBS Open Bio*. https://doi.org/10.1002/2211-5463.13921

Sanusi, I. T., Agbo, F. J., Dada, O. A., Yunusa, A. A., Aruleba, K. D., Obaido, G., Olawumi, O., Oyelere, S. S., & Centre for

Multidisciplinary Research and Innovation (CEMRI). (2024). Stakeholders' insights on artificial intelligence education: Perspectives of teachers, students, and policymakers. *Computers and Education Open*, *7*, 100212. https://doi.org/10.1016/j.caeo.2024.100212

Shabatura, J. (2022). *Using Bloom's Taxonomy to write effective learning outcomes*. University of Arkansas. https://tips.uark.edu/using-blooms-taxonomy/#gsc.tab=0

Siegesmund, A. (2017). Using self-assessment to develop metacognition and self-regulated learners. *FEMS Microbiology Letters*, *364*(11). https://doi.org/10.1093/femsle/fnx096

Siemens, G., Marmolejo-Ramos, F., Gabriel, F., Medeiros, K., Marrone, R., Joksimovic, S., & de Laat, M. (2022). Human and artificial cognition. *Computers and Education: Artificial Intelligence*, *3*, 100107. https://doi.org/10.1016/j.caeai.2022.100107

Stancic, M. (2020). Peer assessment as a learning and self-assessment tool: A look inside the black box. *Assessment & Evaluation in Higher Education*, *46*. https://doi.org/10.1080/02602938.2020.1828267

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

The University of Melbourne. (2023a, July 27). *Design nested or staged assessments*. Assessment, AI and Academic Integrity. Retrieved December 13, 2024, from https://melbourne-cshe.unimelb.edu.au/ai-aai/home/ai-assessment/designing-assessment-tasks-that-are-less-vulnerable-to-ai/seven-practical-strategies/3.-design-nested-or-staged-assessments

The University of Melbourne. (2023b, July 27). *Incorporate more authentic, context-specific, or personal assignments*. Assessment, AI and Academic Integrity. Retrieved December 13, 2024, from https://melbourne-cshe.unimelb.edu.au/ai-aai/home/ai-assessment/designing-assessment-tasks-that-are-less-vulnerable-to-ai/seven-practical-strategies/5.-incorporate-more-authentic,-context-specific,-or-personal-assignments

The University of Melbourne. (2023c, July 27). *Incorporate tasks that ask students to demonstrate evaluative judgement*. Assessment, AI and Academic Integrity. Retrieved December 13, 2024, from https://melbourne-cshe.unimelb.edu.au/ai-aai/home/ai-assessment/designing-assessment-tasks-that-are-less-vulnerable-to-ai/seven-practical-strategies/2.-incorporate-tasks-that-ask-students-to-demonstrate-evaluative-judgement

Thea, W. (2021, January 20). *Designing and implementing learning outcomes*. The University of Sydney. https://educational-innovation.sydney.edu.au/teaching@sydney/designing-and-implementing-learning-outcomes/

Thompson, J. D. (2023, February 3). *Don't panic: There are better ways for universities to respond to ChatGPT*. ABC Religion & Ethics. https://www.abc.net.au/religion/why-universities-should-stop-panicking-about-chatgpt/101929400

Timperley, H. (2020). Using assessment data for improving teaching practice. *Professional Educator*, *8*(3), 28–31. https://doi.org/10.3316/aeipt.179743

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region*.

Tomas, C., & Jessop, T. (2019). Struggling and juggling: A comparison of student assessment loads across research and teaching-intensive universities. *Assessment & Evaluation in Higher Education, 44* (1), 1-10. https://doi.org/10.1080/02602938.2018.1463355

Treleaven, L., & Voola, R. (2008). Integrating the development of graduate attributes through constructive alignment. *Journal of Marketing Education*, *30*, 160–173. https://doi.org/10.1177/0273475308319352

UNSW. (n.d.). *Guidance on AI in assessment*. Retrieved December 6, 2024, from https://www.teaching.unsw.edu.au/ai/ai-assessment-guidance

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: The crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education*, *43*, 1–16. https://doi.org/10.1080/02602938.2018.1427698

Weng, J. (2023). *putting intellectual robots to work: Implementing generative AI tools in project management: Faculty digital archive: NYU libraries*. NYU SPS Applied Analytics Laboratory. https://archive.nyu.edu/handle/2451/69531

Winter, P. C., Kopriva, R. J., Chen, Ch S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences, 16*, 267–276. https://doi.org/10.1016/j.lindif.2007.01.001

Zainudden, D., Broom, M., Nousek-McGregor, A., Stubbs, F., & Veitch, N. (2022). Embedding 21st century employability into assessment and feedback practice through a student–staff partnership. *Access Microbiology*, *4*(3), 000329. https://doi.org/10.1099/acmi.0.000329