Vol.8 No.2 (2025)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

How (not) to use the AI Assessment Scale

Mike Perkins[A] A Centre for Research & Innovation, British University Vietnam, Vietnam

Jasper Roe[B] B School of Education, Durham University, United Kingdom

Leon Furze[C] C School of Education, Deakin University, Australia

DOI: https://doi.org/10.37074/jalt.2025.8.2.15

Abstract

The rapid uptake of generative AI (GenAI) has exposed weaknesses in assessment design and policy. The AI Assessment Scale (AIAS) offers a practical way to align permitted AI use with intended learning outcomes. However, as the scale has grown, we have seen implementations that we feel could be improved: unenforceable 'No AI' labels, assigning labels to existing assessments without changing the assessments themselves, equity blind spots, and control-first policies that encourage 'performance theatre'.

This commentary outlines our recommendations for AIAS implementations that replace detection with design: selecting levels by outcome and conditions, requiring light process evidence where suitable, adjusting criteria to assess judgement and voice, sequencing evidence across time, and planning for equitable access. We highlight sector-specific considerations for implementing and adapting the AIAS for K-12, higher education, Technical and Vocational Education and Training (TVET), English as a Foreign Language (EFL), and English for Academic Purposes (EAP).

Keywords: AI Assessment Scale (AIAS); assessment design; educational assessment; Generative AI; validity.

Introduction

Educators need a clear way to design valid assessments in a GenAI-saturated world, and one way this has been achieved is through the AI Assessment Scale (AIAS) (Perkins, Furze, et al., 2024; Perkins, Roe, & Furze, 2024). This commentary does three things. First, it describes the most common ways in which the AIAS has been unintentionally misused or applied. Second, we suggest how the AIAS can be effectively applied to support GenAI integration into assessments. Finally, we offer suggestions as to how this might be done in K-12, higher education (HE), and Technical and Vocational Education and Training (TVET), so that readers can adapt the approach to their context.

The AIAS began as a practical response to a specific moment. In early 2023, Australian media reported that the Group of Eight universities were reverting to pen-and-paper examinations to counter the use of GenAI tools such as ChatGPT by students in open assessments (Cassidy, 2023). At this point, the authors recognised the need for an alternative approach to assessment that was more sustainable and equitable than reverting entirely to end of unit examinations. With this in mind, Leon Furze drafted an initial framework to move discussions away from blanket bans and toward design choices recognising the possibility of GenAI integration into assessment practices (Furze, 2023). Through collaboration with Mike Perkins, Jasper Roe, and Jason MacVaugh, we refined the early sketch into a more structured model and circulated a preprint of the AIAS in December 2023 to help educators act during what felt like an assessment emergency (Perkins et al., 2023).

By the time the AIAS was published in the *Journal of University Learning & Teaching Practice* in April 2024 (Perkins, Furze, et al., 2024), the assessment landscape had already shifted. Evidence has been accumulating about the fallibility of AI text detectors, with significant false positives and bias risks that make them unsuitable for high-stakes decisions (Chaka, 2023, 2024; Perkins, Roe, Postma, et al., 2024; Perkins, Roe, Vu, et al., 2024; Sadasivan et al., 2024; Weber-Wulff et al., 2023). Our original version of the scale implied the use of detector checks as a control mechanism at earlier levels, which we recognised was neither a reliable nor an educationally useful approach for summative assessment. We concluded that the scale should not depend on technologies that cannot accurately demonstrate who wrote what, and that further adjustment was needed. We recognised growing calls for assessment redesign in the GenAI era, moving away from concepts of detection and policing, and more towards supporting instructors to develop valid assessments for the

age of GenAI (Bearman et al., 2024; Lodge, Howard, et al., 2023). These concerns sit within longer-standing plagiarism typologies that predate GenAI, including complex forms of plagiarism that evade simple detection (Perkins et al., 2019). Recognising that having valid assessments may matter more than concerns of student cheating (Dawson et al., 2024), especially in the uncertain times that the education sector was facing, we drew on a Vygotskian view of learning (Vygotsky, 1978) to redevelop the AIAS. Using the perspective of GenAI as a cultural tool that may mediate knowledge construction when used by students, as a scaffold to help students achieve more than they would otherwise be able to do within their Zone of Proximal Development (ZPD), we were able to separate two questions that are often conflated. First, what do we want students to be able to do unaided here and now? Second, how do we want them to use and evaluate GenAI tools as part of their authentic practice?

We refocused on developing the AIAS to do both, and when considered from this perspective, it avoids the trap of being a purely discursive approach (Corbin, Dawson, et al., 2025) by giving staff and students a shared language for permitted use in a given task and acting as a scaffold for educators to redevelop and redesign assessments. This rebalancing toward judgement and critique aligns with evidence that GenAI can support, but should not replace, learner agency (Roe & Perkins, 2024a). From this perspective, we adjusted the model, moving to five non-hierarchical levels: 'No AI' (Level 1), 'AI Planning' (Level 2), 'AI Collaboration' (Level 3), 'Full AI' (Level 4), and 'AI Exploration' (Level 5). The updated AIAS is shown in Figure 1.

The AI Assessment Scale

1	NO AI	The assessment is completed entirely without AI assistance in a controlled environment, ensuring that students rely solely on their existing knowledge, understanding, and skills **You must not use AI at any point during the assessment. You must demonstrate your core skills and knowledge.**
2	AI PLANNING	AI may be used for pre-task activities such as brainstorming, outlining and initial research. This level focuses on the effective use of AI for planning, synthesis, and ideation, but assessments should emphasise the ability to develop and refine these ideas independently. **You may use AI for planning, idea development, and research. Your final submission should show how you have developed and refined these ideas.**
3	AI COLLABORATION	AI may be used to help complete the task, including idea generation, drafting, feedback, and refinement. Students should critically evaluate and modify the AI suggested outputs, demonstrating their understanding. **You may use AI to assist with specific tasks such as drafting text, refining and evaluating your work. You must critically evaluate and modify any AI-generated content you use.**
4	FULL AI	AI may be used to complete any elements of the task, with students directing AI to achieve the assessment goals. Assessments at this level may also require engagement with AI to achieve goals and solve problems. **You may use AI extensively throughout your work either as you wish, or as specifically directed in your assessment. Focus on directing AI to achieve your goals while demonstrating your critical thinking.**
5	AI EXPLORATION	AI is used creatively to enhance problem-solving, generate novel insights, or develop innovative solutions to solve problems. Students and educators co-design assessments to explore unique AI applications within the field of study. **You should use AI creatively to solve the task, potentially co-designing new approaches with your instructor.**

Figure 1. The Updated AIAS.

The AIAS has been widely adopted in K-12 (Furze, Perkins, & Roe, 2024) and HE (Furze, Perkins, Roe, et al., 2024) contexts, as well as in TVET (Riis, 2025). Ongoing research is being carried out across the globe on the potential that this has to support learners and educators in adapting to this 'wicked problem' (Corbin, Bearman, et al., 2025). To date, we are aware of more than 350 institutions using the scale in one form or another. It has been translated into over 30 languages, cited in policy and course documentation across multiple systems, and has received recognition from organisations around the globe as a method of supporting academic integrity (Studiosity, 2024, 2025) and noted by educational regulators such as the Australian Tertiary Education Quality and Standards Agency as an option to assist with implementing GenAI into assessment (Lodge, 2024).

This commentary builds on our experiences discussing AIAS implementations with educators and administrators worldwide by describing some of the most common ways in which we have seen the AIAS used in ways that do not necessarily support assessment validity or task redesign. The aim of this is not to criticise; in our view, doing something and proactively taking control is better than doing nothing at all. Early users of the AIAS have contributed to the development of the discourse surrounding GenAI in assessment and provided direction for refining the scale. Those institutions that tried to do something, rather than ignoring the challenges of GenAI, have acted as trailblazers for other, perhaps more cautious institutions to follow in their footsteps. Our purpose is to help others learn from their actions and outline what effective implementation looks like when the scale is used as both a redesign framework and communication device.

Second, we offer suggestions and advice based on our collective work with institutions worldwide to guide future implementations of the AIAS and highlight what should be considered before doing so. Finally, we discuss how AIAS implementations may need to be adapted when carried out across a variety of contexts, including K-12, HE, TVET, and in English as a foreign language (EFL) and English for academic practice (EAP). All these contexts have different needs, regulatory environments, and unique challenges for assessment which need to be addressed in unique ways.

In all cases, the fundamental challenge for implementing GenAI in assessment remains the same: if the task and its conditions do not change despite the introduction of GenAI as a tool which can be used by students (illicitly or not), then challenges with assessment validity will occur. We hope to demonstrate how institutions can avoid integrating GenAI into assessment without it becoming either an example of 'security theatre' which is swiftly ignored by students, leading to academic integrity concerns, or worse, pretending that students can demonstrate their mastery of learning outcomes just as well with GenAI compared to without it, leading to a false sense of knowledge creation.

How *not* to use the AIAS

One of the key affordances of the AIAS is its flexibility and ability to be tailored to meet a wide range of assessment tasks in multiple educational contexts. That said, there are several areas that users should be aware of when seeking to implement the AIAS, which have mainly come to light since the original release of the first iteration of the scale. In this section, we reflect on some of the ways that the AIAS could be used inappropriately or in a way that does not support effective, valid assessment.

When can it really be 'No AI'?

The first major misconception is that the AIAS can be used to label an existing assessment without considering the possibility of enforcement. For example, we have seen in some cases educators labelling take-home essays as 'No AI'. In an unsecured essay assessment, there is no realistic way to ensure that no AI technologies are used throughout the process. It is understandable that some educators will wish to take this approach; after all, many conversations we have had suggest that educators wish they could go back to the 'old days' when they only had to worry about contract cheating, collusion, or other forms of misconduct, rather than wrap their heads around what AI means for these concepts. This is impossible because there are no accurate ways to detect the use of AI in written work with current technologies, and it seems unlikely that any will emerge in the future. For every detection technology, there will quickly arise a way around it, whether that is high tech (a new model which is less detectable, as we saw from GPT-3 to GPT-4) or low tech (copying and pasting different passages, manipulating AI text to sound more 'human'). This is something we have written about prior to the public release of ChatGPT or other GenAI models. In 2022, we claimed that Automated Paraphrasing Tool (APT) detection and evasion constituted an arms race of development and evasion which ultimately is a zero-sum game and benefits neither teachers nor learners (Roe & Perkins, 2022).

Using this kind of approach ultimately creates conditions that are purely performative: students may declare 'No AI', teachers may believe 'No AI', yet the assessment is both invalid and insecure. Operating assessments in this way, with no adjustment to structural changes, has been referred to as a 'discursive' approach to assessment redesign in the AI era (Corbin, Dawson, et al., 2025) and is not an option we recommend.

Ultimately, we feel that programmes of study should often (and in some cases must) contain 'No AI' assessments, but that these need to be practical and pragmatic. Dawson (2020) states that assessment security must contain authentication (an assurance that the student is the person who completed the work and not someone else) and control of circumstances (the assessment must be taken under specific conditions). This does not mean that 'No AI' assessments must only be proctored examinations or oral viva voces, however. There are multiple other options in which these conditions of assurance and control of circumstances can be met; for example, in-class discussions, debates, or informal moments

of technology-free interactions are all assessable materials. Where higher-order outcomes are the target, it is more credible to surface student decision-making about tool use than to attempt blanket prohibition (Roe & Perkins, 2024b).

AI for AI's sake

By that, we mean that an assessment is suddenly upgraded to 'Full AI' without due consideration for what this means for students being able to demonstrate the learning objectives of the task.

A change introducing GenAI technologies fundamentally alters the conditions, outcomes, and meaning of any assessment, and doing this without considering the resulting changes in learning does no favours to the learner or the assessor. Changing assessments to 'Full AI' or 'AI Exploration' needs to involve a major redesign of the task to account for the assistance provided by GenAI tools and not just be a way to avoid engaging with the concept of academic dishonesty or inappropriate use of AI by students. We believe it is better to engage with AI and enable or advocate for its use where it adds value in a way that has external validity, that is, it is how AI would be used in a target domain outside of the programme of education, rather than adding AI for AI's sake.

Retrofit without redesign

Both issues fall into the trap of attempting to retrofit AIAS levels to the assessment task without considering the context in which it is being carried out. We are aware that assessment design is difficult, time-consuming, and requires significant bureaucratic considerations in some contexts (i.e. plans for assessment needing to be submitted several months or even years in advance). In these cases, it might be tempting for educators to just assign an AIAS level ('Okay, you can use Level 3 in this task...') without making any changes to the assessment brief, rubric, or student guidance. In relation to this, we see the AIAS primarily as an assessment *design* tool, not an assessment security tool. For the reasons explained elsewhere in this commentary, it is not possible or fair to simply label a task at a given level of the AIAS and then hope that students will comply with it. If you permit GenAI, redesign the brief, evidence trail, and rubric to grade students' decisions, checks, and justifications, not just their ability to push buttons.

Equity and inclusion

AI use is divisive; some educators seem to embrace it fully and look toward a near-future in which AI is part and parcel of our everyday lives (for many of us, it already is, whether we know it or not). Others actively resist the incursion of AI and its use in education on moral, ethical, environmental, and other grounds. These concerns are valid, but we do not see active AI resistance as a tenable option as long as AI tools are publicly accessible to learners. The reason that this matters, and we advocate for using the AIAS (even if you are an AI resistor), is to ensure that assessment is valid and fair, but also to ensure equity as far as is possible.

Ignoring GenAI risks inequity and invalid results. Just as some educators are heavier users of AI, some students are more willing, capable, and able to experiment and pay for premium or advanced AI tools. One way that we can address this is to try and ensure that if we design a 'Full AI' assessment, we are prepared for the fact that some learners may be better resourced than others if we are not providing access to the most up-to-date technology (a student with paid access to the very latest model by OpenAI, Claude, or Gemini may have a significant advantage compared to a student using the equivalent free model). Likewise, if we are discursively claiming that a take-home task is 'No AI', and relying on text detection technology to enforce this, those students who have access to the most advanced models and are able to use AI humanising tools to manipulate the resulting output will go undetected, gaining an unfair advantage over their peers.

Furthermore, we must consider the overall assessment approaches taken in the context of the module or programme. As mentioned, some educators take different stances on AI use in their teaching and assessment strategies. If a module is changed from year to year with a different instructor, perhaps from one who allows AI use in assessment to one who does not, then this negatively affects score validity and creates fairness issues across programmes. This effect may be multiplied when this happens on a section-by-section basis; we feel strongly that AI assessment policies should be standardised as much as possible within one module or unit to ensure that fairness and equity are maintained.

Flexible not fixed

We are grateful for the support, interest, and debate generated by the AIAS. At the time of the first release, it was the best option we could offer to deal with the disruption GenAI created in our professions. Since then, new approaches to integrating GenAI in assessments, such as the two-lane approach (Liu & Bridgeman, 2023), new technologies such as agentic AI, and shifts in our thinking have shaped its evolution. In response, we refined and developed the framework collaboratively with input from many other scholars.

We do not claim to know every context. This is why we published the AIAS under a Creative Commons licence and encouraged adaptation, remixing, and alteration of the scale to suit different contexts and assessment types. As a result, we have seen many different ways of utilising the AIAS across programmes, levels of education, and cultural contexts. Some of these seem to us more intuitive and practical than others, but we are always supportive of any attempts to use the AIAS to try and ensure equitable and valid assessment. If the current form does not fit, keep the principle of redesigning assessment for a new GenAI reality and create something that works for you.

For us, the levels of the AIAS are not prescriptive; they are guiding. For example, we frequently receive questions asking specifically what the difference is between a Level 3 and a Level 4 assessment task, or between a Level 4 and a Level 5 task. However, we believe that the interpretation

of this is up to individual educators, and finer distinctions can be co-defined by the educator, institution, and learner. We do not see the levels of the AIAS as a rigid structure that must be adhered to, but as ways of thinking about how assessment can be reconfigured to support the assessment of and for learning that can be adapted and altered to suit the on-the-ground conditions.

Implementation suggestions

This section offers general tips and advice based on our collective work with schools, universities, businesses, and education providers worldwide. As with the rest of this commentary, it is not offered as a hard-and-fast set of rules, but as examples of things that we have found work in practice to support effective assessment, as well as things to avoid.

Focus on the faculties

We have found that working at the subject, discipline, and faculty levels is the most successful approach to implementing the AIAS in an institution. 'Leading from the middle' is an effective way to disseminate training and professional learning in schools (Lipscombe et al., 2023) and makes sense when working with GenAI as it is highly contextual for each discipline: what GenAI tools mean to assessments set by an English teacher is very different to the mathematical reasoning implications of the same model used in a Maths or Science classroom.
This is important because each field has unique relationships with how text production, problem-solving, and creative outputs are valued and used, and these discipline-specific needs cannot be addressed through generic, institution-wide policies alone. Working at the faculty level with groups of subject-matter experts allows for the development of assessments that can maintain disciplinary integrity in technology integration.

Before using the AIAS

Before faculties begin to work with the AIAS, we recommend an honest discussion about the overall validity of their current assessments. In many contexts, particularly senior secondary and HE, there has been a drift towards high-stakes examination-style assessments for most or all formally graded outcomes as GenAI usage has become more widespread. We consider this approach to lack validity because these 'secure' assessments are not inclusive or accessible (Dawson, 2022). Frank conversations in faculty meetings about whether existing assessments are valid should precede any consideration of GenAI's implications.

Content validity (are we really assessing what we have taught?) and construct validity (does this assessment measure the intended knowledge and skills?) are also important considerations before worrying about GenAI. We need to be clear on whether our existing assessments achieve the outcomes we are looking for. Again, examinations are a good example: while they are a secure Level 1 type of assessment, they may not be the best way to judge a student's capabilities in all topics or disciplines.

Determining the role of GenAI in learning

When considering GenAI integration, educators should evaluate whether the technology being discussed helps or harms the learning objectives. The key question becomes: 'Is it a bad use of the technology, or a bad use of your brain?' If the answer to either of these questions is '' yes, then there is a strong chance that a Level 1 task might be necessary. Otherwise, Levels 2-5 of the AIAS can be explored as options. Faculty discussions should address whether the skills or knowledge being assessed might be too easily offloaded to GenAI, and what mitigation strategies might be needed. However, automatic assumptions that all AI use by students in assessments constitutes 'cheating' should be avoided.

While certain skills may be diminished by GenAI use, we urge faculties to consider which skills might be 'let go'. Historically, we have seen this happen many times: GPS replacing map-reading skills, the calculator offsetting the need for mental arithmetic (Lodge, Yang, et al., 2023). GenAI will potentially lead to some skill loss, and it is too early to determine exactly what that will look like. As disciplinary experts, educators must judge which subject-specific skills and knowledge need to be retained and which can, in some ways, be supported by AI tools.

Choosing the right level for each task

Finally, selecting an appropriate level for a task is important. We say 'task' because summative assessment should be broken down into multiple points of assessment over time to support validity, rather than being conducted as a single high-stakes end-of-unit test. (Dawson et al., 2024; van der Vleuten et al., 2012). At the start of a unit, it might make sense to have an informal Level 1 task, such as an observed group discussion, to benchmark students' initial understanding of a topic. Research tasks can be conducted using GenAI 'Deep Research' tools (Level 2). A Level 1 draft written in class might be used as a link in the evidence chain to authenticate a later (Level 3) GenAI-edited final version. In some cases, Level 4 or 5 might be the best choice because an open and unrestricted use of AI best reflects the reality of how students will be expected to interact with the technology outside of education. Equity is a key consideration here: if a task *requires* students to use GenAI tools to complete it, guarantee access to free tools. If Level 1 is used, accommodations for any assistive technology that involves AI must be explicit, distinguishing it from AI that would compromise validity.

Overall, we support educators' autonomy and professionalism in deciding where GenAI should and should not be used. These decisions must be made with eyes wide open about the technology, not on assumptions or personal preference, and they should be made with students in open, transparent conversations that explain why the level was chosen, what AI uses are permitted or prohibited, what evidence and disclosure are required, and what access and

accommodations are in place.

Contextual variations matter

The AIAS authors teach across diverse contexts, including HE, K-12, adult education, TVET, and EAP/ EFL. The AIAS has also been applied in business settings for corporate learning and development purposes. Given the vast differences between educational contexts across jurisdictions, the AIAS requires a flexible application.

Since we first published the original AIAS, we have created permissively licenced CC-BY-NC-SA 4.0 documents available on aiassessmentscale.com and published in open-access journals, an approach which has allowed the broad education community to adapt and remix the ideas of the AIAS. The following sections describe how we suggest the AIAS could be applied in the different sectors we work in, both through our teaching and through the advisory and professional development services we provide to schools, universities and training providers.

Higher education

The variation in assessment control mechanisms across international and institutional contexts in HE is a significant factor in how any implementation of the AIAS should be carried out. Some universities may mandate multiple review steps before assessments are released to students, while others conduct minimal checks and leave considerable flexibility for individual faculty members. In settings with high amounts of flexibility and limited centralised restrictions (which is more common in North American contexts), rapid experimentation is enabled when desired by educators, but it can also cause challenges for consistent student experiences and understanding across a programme, leading to fairness and equity concerns. One way to support faculty in these contexts is to work in a 'leading through the middle' approach, as discussed earlier. This can be achieved by recruiting 'discipline champions' who can support others to ideate tasks and then share these as exemplars with others in their discipline. In these settings which may not include standardised assessment templates, there will still likely be wide variation; therefore, any quality assurance efforts should focus on ensuring that the task mechanics match the declared level.

In more centralised systems, such as the United Kingdom, we believe that a combination of top-down policy and bottom-up design works well. Programme or discipline leaders might mandate the student-facing rationale and the requirements for any process evidence, while subject teams choose the levels and task types that fit disciplinary outcomes. Regardless of the level of centralised control, staff development focused on developing and teaching critical AI literacy (Roe, Furze, & Perkins, 2025a, 2025b) is important to support an understanding of the concepts and technologies under discussion, as well as the ability to pass this on to students. Further training focused on the practical skills of assessment redesign and knowledge of the AIAS is also important; we cannot expect all faculty to implement top-down diktats mandating the use of any assessment framework if no support is provided. At the HE level, there is also more scope to design authentic, domain-specific tasks that reward critique, synthesis, creativity, and orchestration rather than mere recall. This is consistent with current work that frames evaluative judgement as a central graduate capability in a time of GenAI (Bearman et al., 2024).

Ultimately, although institutional contexts and control mechanisms vary significantly across HE globally, the AIAS provides a flexible framework that can be adapted to local conditions while maintaining focus on the core objective of ensuring students develop both disciplinary expertise and the critical capabilities needed to work effectively with GenAI in their future professional lives. This mirrors current developments in the application of GenAI to research practice, where GenAI is becoming more integrated throughout the entire research cycle (Bjelobaba et al., 2025; Perkins & Roe, 2024a, 2024b), making transparency and justification of GenAI use core graduate capabilities.

K-12

K-12 education, particularly senior secondary assessment, operates within the framework of high-stakes standardised testing in most jurisdictions. This context has complicated discourses about AI, as examinations often become the 'tail that wags the dog' for many teachers and schools, that is, students must do an exam, and therefore all or most assessments in years 7-12 are turned into miniature exams. AI has been used as an excuse to move towards more end-of-unit, supervised, timed examinations, since those assessments are seen as both 'more secure' and reflect the formal exams that often come at the end of secondary schooling. However, this approach contradicts the best practices in assessment design. Valid assessments must be equitable, authentic, replicate real-world and practical skills, vary in form and mode, and enable teachers to build comprehensive pictures of students over time. While examinations are secure, they fail to meet these criteria (Dawson et al., 2024). Following Dawson et al.'s (2024) argument that AI use becomes unacceptable when validity is threatened, valid high-stakes assessments prohibiting AI become impossible if the prohibition cannot be enforced. This shifts the conversation from 'cheating' to adjusting assessment types and conditions.

The AIAS in K-12 contexts helps mediate this discourse by offering approaches to responsibly involve GenAI without defaulting to examinations. The framework supports validity in assessment through explicit teaching of outcomes outlined in curricula (content validity), selecting optimal assessment modes for outcomes (construct validity), considering design consequences on student behaviour (consequential validity), creating authentic assessments reflecting real-world applications, designing varied assessment formats, ensuring inclusivity and accessibility, combining formal and informal assessment while valuing professional judgements alongside external validation, and building assessment evidence over time to develop comprehensive student profiles. The 'Swiss Cheese' approach in education (Morley & Zmood, 2015; Rundle et al., 2020) provides a useful way

of combining these tasks to gain a greater overall awareness of the true performance of a student: assessments can be broken into separate tasks, some formal and others informal, with different 'slices' set at various AIAS levels. By building a unit of work with multiple points of assessment aligned to different levels of the scale, we can support students to have both opportunities to use the technology in authentic ways and methods of assessment that demonstrate their 'AI-free' knowledge and skills at other points. The chain of evidence is built up over time, and the teacher is not reliant on a single, high-stakes examination as the measure of a student's ability to attain the required learning outcomes.

Technical and vocational education and training

TVET demonstrates strong positioning for GenAI benefits, as assessment methods already largely exemplify the best practices outlined above. Vocational studies, in our experience, tend to focus more on observable skills and are much more likely than senior secondary or even HE to include assessments such as simulation, role play, direct observation, on-the-job Q&A, and scenario-based tasks. All of these assessment styles are less vulnerable to the kind of low-stakes academic integrity violations encountered by teachers whose primary mode of assessment is the essay or short/long answer written questions.

Because vocational assessment methods vary over time and typically lead to satisfactory/not satisfactory or competent/not competent marks rather than numerical grades, they appear less susceptible to cheating behaviours encountered in higher education settings, although we recognise that even practical assessment tasks remain vulnerable to cheating. Riis (2025) offers a TVET-specific articulation of GenAI integration inspired by the AIAS: five strategies from AI independent learning to AI-centric learning, illustrated through carpentry and grounded in socio-material perspectives. Together, these developments point to a coherent path for valid assessment in skills-based programmes

Many vocational courses have literacy and numeracy requirements, and we believe that AI can be used intelligently in these scenarios. If literacy assessment primarily concerns communication, Level 3 tasks that allow extensive GenAI use for drafting and editing written communications may be appropriate. Level 2 approaches might involve verbal note-taking, recording, transcription, and AI refinement, rather than starting from blank pages. Level 1 assessments in TVET are potentially easier to manage. At the time of writing, ChatGPT is proficient in writing essays but cannot rewire an electrical circuit, insert an intravenous drip, fix a cistern, or support a child in early childhood education with additional needs. All of these are observable skills that would be best assessed in a 'No AI', technology-free environment.

English as a foreign language and English for academic purposes

English as a Foreign Language (EFL) and English for Academic Purposes (EAP) contexts present a specific problem: when writing is the learning outcome, full text generation obscures whether learners have achieved the target construct (i.e. writing). Based on the authors' experience in these contexts and in the field of English as a Medium of Instruction (EMI) for EFL speakers, we developed specific adaptations for the AIAS. Our adaptations of the AIAS respond to this by defining alternate levels of the scale which protect evidence of independent language production while still teaching AI literacy (Roe et al., 2024; Roe, Furze, Perkins, et al., 2025; Roe, Perkins, et al., 2025) and encouraging the use of GenAI to potentially assist in language learning.

In practice, we recommend using Level 1 tasks for core writing and speaking under controlled conditions. Where these Level 1 tasks are written, we would encourage a short oral check to triangulate proficiency. Level 2 might be used for broader EAP tasks for planning and idea generation, with a brief rationale that shows how AI-sourced ideas were selected, refined, or rejected (note that EAP-style tasks can be used for both English as a first language and EFL speakers). When Level 3 is used for editing support, we recommend choosing a method that leaves a visible trail of changes, for example, a 'clean' final version plus a tracked-changes version with a short commentary. Higher levels of the AIAS should be reserved for assessing the production of specified communication pieces rather than unaided language production, such as genre adaptation for a specific audience. Rubrics should make this split explicit by separating language proficiency from content knowledge and AI-mediated communication.

This approach aligns with a broader point about learner agency. Students studying in a second language are often described as passive or lacking autonomy. Evidence shows that these are misconceptions. When provided with clear guidance, appropriate tools, and structured opportunities, these students readily engage in autonomous learning practices and can make sophisticated choices about when and how to use support, including AI (Roe & Perkins, 2020). The task is to design assessment conditions that allow students to demonstrate, rather than hide, this competency. The task is to design assessment conditions that allow students to make their agency visible by directing the tools, showing their checks, and accounting for access constraints (Roe & Perkins, 2024a).

Conclusion

Assessment in the GenAI era has changed dramatically and quickly, but our overall goals have not changed. We still need trustworthy evidence of what students know and can do, gathered under conditions that make sense for the outcome of the assessment. The AIAS can help when it is used to make these conditions explicit and redesign tasks. In this commentary, we mapped the most common challenges in implementing the AIAS and demonstrated how they undermine assessment validity.

Our experience shows three pitfalls that erode validity: declaring 'No AI' where enforcement is impossible, adding AI for its own sake, and retrofitting levels without redesigning tasks, rubrics, and student guidance. Equity is a parallel risk

for broader institutional trust if GenAI access, consistency, and accommodations are not planned across a programme or institution. We have illustrated how implementations differ across HE, K-12, TVET, and EFL/EAP, but the key challenges remain the same across all contexts.

Our key suggestions for the effective implementation of the AIAS are as follows:

1. Audit the broader validity of the assessments currently used;
2. Decide the appropriate AIAS level per task, then redesign the brief, evidence trail, and rubric to fit that choice;
3. Communicate permitted and prohibited uses in plain language and back it up through structural redesign;
4. Build a chain of evidence of student attainment over time, rather than relying on single high-stakes moments;
5. Align approaches within faculties to respect disciplinary and institutional norms while ensuring consistency for students;
6. Build faculty capability through training that includes supporting critical AI literacy and principles of assessment design;
7. Recognise equity issues and guarantee access to the required tools for all students.

The AIAS is not intended as a tool to optimistically laud the integration of AI into education, and there is good reason to resist AI in multiple educational contexts; critical analysis and scepticism should be our default position. However, the AIAS provides a framework within which to consider technological change, update our approach to assessing students, and offer transparency and clarity to all educational stakeholders when used in line with the principles outlined above. By integrating GenAI into assessment in this way, we can take steps to reduce broader structural inequities in GenAI assessment integration: redesign for what matters and be honest about what AI changes and what it does not (Perkins & Roe, 2025).

Acknowledgements

We thank all educators worldwide who have adapted, implemented, commented on, or otherwise engaged with all versions of the AIAS. Without your engagement, we would not be able to offer the guidance included in this commentary for future users of the AIAS.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors have used GenAI tools for ideation and in some passages of draft text creation which was then heavily revised, along with editing and revision during the production of the manuscript. The tools used were ChatGPT (GPT-5) and Claude (4.1 Opus), which were chosen for their ability to provide sophisticated feedback on textual outputs. These tools were selected and used supportively and not to replace core author responsibilities and activities. The authors reviewed, edited, and take responsibility for all outputs of the tools used.

References

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, 0*(0), 1–13. https://doi.org/10.1080/02602938.2024.2335321

Bjelobaba, S., Waddington, L., Perkins, M., Foltýnek, T., Bhattacharyya, S., & Weber-Wulff, D. (2025). Maintaining research integrity in the age of GenAI: An analysis of ethical challenges and recommendations to researchers. *International Journal for Educational Integrity, 21*(1), Article 1. https://doi.org/10.1007/s40979-025-00191-w

Cassidy, C. (2023, January 10). Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. *The Guardian*. https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2), 94-104. https://doi.org/10.37074/jalt.2023.6.2.12

Chaka, C. (2024). Accuracy pecking order – How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning and Teaching, 7*(1), 127-139. https://doi.org/10.37074/jalt.2024.7.1.33

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education, 0*(0), 1–17. https://doi.org/10.1080/02602938.2025.2553340

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education, 0*(0), 1–11. https://doi.org/10.1080/02602938.2025.2503964

Dawson, P. (2020). Assessment security. In *Defending assessment security in a digital world*. Routledge.

Dawson, P. (2022). Inclusion, cheating, and academic

integrity: Validity as a goal and a mediating concept. In *Assessment for Inclusion in Higher Education*. Routledge.

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 0*(0), 1–12. https://doi.org/10.1080/02602938.2024.2386662

Furze, L. (2023). *The AI Assessment Scale: From no AI to full AI – Leon Furze* [Blog]. https://leonfurze.com/2023/04/29/the-ai-assessment-scale-from-no-ai-to-full-ai/

Furze, L., Perkins, M., & Roe, J. (2024). The AI Assessment Scale (AIAS) in Australian K–12 education. *Teachers' Frontiers, 1*(1), 17–24. https://www.researchgate.net/publication/386347227_The_AI_Assessment_Scale_AIAS_in_Australian_K-12_Education

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*. https://doi.org/10.14742/ajet.9434

Lipscombe, K., Buckley-Walker, K., & Tindall-Ford, S. (2023). Middle leaders' facilitation of teacher learning in collaborative teams. *School Leadership & Management, 43*(3), 301–321. https://doi.org/10.1080/13632434.2023.2215803

Liu, D., & Bridgeman, A. (2023, July 12). *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. M. (2024). *The evolving risk to academic integrity posed by generative artificial intelligence: Options for immediate action* | Tertiary Education Quality and Standards Agency. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/evolving-risk-academic-integrity-posed-generative-artificial-intelligence-options-immediate-action

Lodge, J. M., Howard, S., Bearman, M., & Dawson, P. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence

Lodge, J. M., Yang, S., Furze, L., & Dawson, P. (2023). It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? *Learning: Research and Practice, 9*(2), 117–124. https://doi.org/10.1080/23735082.2023.2261106

Morley, P., & Zmood, S. (2015). Determining a student's optimal learning zone in light of the Swiss Cheese Model. *Mathematics Education Research Group of Australasia*. https://eric.ed.gov/?id=ED572509

Perkins, M., Basar Gezgin, U., & Gordon, R. (2019). Plagiarism in higher education: Classification, causes and controls. *Pan-Pacific Management Science, 2*, 3–21. https://doi.org/10.13140/RG.2.2.20694.11841

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2023). *Navigating the generative AI era: Introducing the AI assessment scale for ethical GenAI assessment* (No. arXiv:2312.07086). arXiv. https://doi.org/10.48550/arXiv.2312.07086

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of Generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(06), Article 06. https://doi.org/10.53761/q3azde36

Perkins, M., & Roe, J. (2024a). Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis [version 2; peer review: 3 approved, 1 approved with reservations]. In *F1000Research* (Vol. 12, Issue 1398). https://doi.org/10.12688/f1000research.142411.2

Perkins, M., & Roe, J. (2024b). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning and Teaching, 7*(1), 390-395. https://doi.org/10.37074/jalt.2024.7.1.22

Perkins, M., & Roe, J. (2025). The end of assessment as we know it: GenAI, inequality and the future of knowing. In UNESCO (Ed.), *AI and the future of education: Disruptions, dilemmas and directions* (pp. 76–80). UNESCO. https://doi.org/10.54675/KECK1261

Perkins, M., Roe, J., & Furze, L. (2024). *The AI Assessment Scale revisited: A framework for educational assessment* (No. arXiv:2412.09029). arXiv. https://doi.org/10.48550/arXiv.2412.09029

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). 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

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

Riis, M. (2025, August 19). *Between wood and algorithms: Crafting AI literacy in TVET* | UNESCO. UNESCO Futures of Education. https://www.unesco.org/en/articles/between-wood-and-algorithms-crafting-ai-literacy-tvet

Roe, J., Furze, L., & Perkins, M. (2025a). Digital plastic: A metaphorical framework for critical AI literacy in the multiliteracies era. *Pedagogies: An International Journal, 0*(0), 1–15. https://doi.org/10.1080/1554480X.2025.2557491

Roe, J., Furze, L., & Perkins, M. (2025b). Reflecting reality, amplifying bias? Using metaphors to teach critical AI literacy. *Journal of Interactive Media in Education,* 1–15. https://doi.

org/10.5334/jime.961

Roe, J., Furze, L., Perkins, M., & Kitson, C. (2025, March 17). *The AI Assessment Scale: A practical framework for TESOL educators in the age of ChatGPT.* TESOL Connections | TESOL International Association. https://www.tesol.org/the-ai-assessment-scale-a-practical-framework-for-tesol-educators-in-the-age-of-chatgpt/

Roe, J., & Perkins, M. (2020). Learner autonomy in the Vietnamese EAP context. *Asian Journal of University Education, 16*(1), 13–21. https://doi.org/10.24191/ajue.v16i1.8490

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

Roe, J., & Perkins, M. (2024a). Generative AI and agency in education: A critical scoping review and thematic analysis (No. arXiv:2411.00631). arXiv. https://doi.org/10.48550/arXiv.2411.00631

Roe, J., & Perkins, M. (2024b). *Generative AI in self-directed learning: A scoping review* (No. arXiv:2411.07677). arXiv. https://doi.org/10.48550/arXiv.2411.07677

Roe, J., Perkins, M., & Furze, L. (2025). *From assessment to practice: Implementing the AIAS framework in EFL teaching and learning* (No. arXiv:2501.00964). arXiv. https://doi.org/10.48550/arXiv.2501.00964

Roe, J., Perkins, M., & Tregubova, Y. (2024). The EAP-AIAS: *Adapting the AI Assessment Scale for English for academic purposes* (No. arXiv:2408.01075). arXiv. https://doi.org/10.48550/arXiv.2408.01075

Rundle, K., Curtis, G., & Clare, J. (2020). Why students choose not to cheat. In T. Bretag (Ed.), *A research agenda for academic integrity* (pp. 100–111). Edward Elgar Publishing. https://doi.org/10.4337/9781789903775.00014

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2024). Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=OOgsAZdFOt

Studiosity. (2024). *The Prof Tracey Bretag prize for academic integrity*. https://www.studiosity.com/traceybretagprize

Studiosity. (2025). *The Prof Tracey Bretag prize for academic integrity.* The 2025 Prof Tracey Bretag Prize for Academic Integrity. https://www.studiosity.com/traceybretagprize

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34*(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes.* Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), Article 1. https://doi.org/10.1007/s40979-023-00146-z