



Rethinking Assessment for Generative Artificial Intelligence

Leon Furze

2025



INTRODUCTION

I've been writing about Generative Artificial Intelligence in education since 2022, when I began my PhD in how these technologies are perceived and used by educators. Prior to that, I was a secondary educator and school leader for fifteen years. My experience ranges from Head of English, to Year Level Coordinator, and Director of Teaching and Learning. I've also worked in Higher Education, mostly in Initial teacher Education courses teaching literacy. Since about 2020, I've been experimenting with digital technologies like OpenAI's GPT, and wondering about its impact on education.

The first version of this eBook was published as a collection of my 2023 blog articles on the topic of rethinking assessment.

This updated version includes all of the original content, plus new articles written in 2024-2025. It is split into two sections: all of the AI Assessment Scale articles first, and then everything else. As the technology develops, I believe it is important to stay flexible and open to new methods of assessment.

I hope you find it as useful as the first edition!

Leon

PS: This collection is a “scrape” of my blog, which means that all of the links are still active; however, it sometimes makes the formatting a bit odd. If anything looks weird or out of place, head to the original blog post.



Leon Furze

www.leonfurze.com

Rethinking Assessment for Generative Artificial Intelligence 2025

Compiled on July 14, 2025

Table of Contents

Part One: The AI Assessment Scale

[The AI Assessment Scale: From no AI to full AI \(2023-04-29\)](#)

[The AI Assessment Scale: Version 1 \(2023-12-18\)](#)

[Can the AI Assessment Scale stop students "cheating" with AI?
\(2024-08-09\)](#)

[Updating the AI Assessment Scale \(2024-08-28\)](#)

[AIAS: Why We've Driven Through the Traffic Lights \(2024-09-02\)](#)

[AI Assessment Scale \(AIAS\) Translations: 2025 Updates \(2025-03-19\)](#)

[How I use the AI Assessment Scale: Part 1 \(2025-06-11\)](#)

Part Two: Rethinking Assessment

[Beyond Cheating: Why the ban and block narrative hides the real threats of ChatGPT in education \(2023-01-21\)](#)

[Using ChatGPT for Conferencing and Feedback \(2023-02-08\)](#)

[Teaching Writing in the Age of AI: Assessment and "Cheating" \(2023-02-18\)](#)

[Secondary School Assessment and Artificial Intelligence \(2023-06-13\)](#)

[Generative AI, plagiarism, and "cheating" \(2023-09-20\)](#)

[Rethinking Assessment for Generative AI: Orals and discussions \(2023-09-27\)](#)

[Rethinking Assessment for Generative AI: Beyond the Essay \(2023-10-04\)](#)

[AI Detection in Education is a Dead End \(2024-04-09\)](#)

[Ditch the Detectors: Six Ways to Rethink Assessment for Generative Artificial Intelligence \(2024-05-03\)](#)

[GenAI Strategy: Attack Your Assessments \(2024-05-13\)](#)

[GenAI Strategy: Bullets then Cannonballs \(2024-05-17\)](#)

[GenAI Strategy: Update your Assessments \(2024-05-21\)](#)

[GenAI Strategy: Evaluate and Communicate \(2024-05-27\)](#)

[GenAI Strategy for Faculty Leaders \(2024-05-01\)](#)

[Don't use GenAI to grade student work \(2024-05-27\)](#)

[Racist, Robotic, and Random: More Thoughts on Generative AI Grading \(2024-06-20\)](#)

[Take-home assessments: AI is not the problem \(2025-06-15\)](#)

You may use AI for planning, idea development, and research. Your final submission should be developed and refined these ideas.

PART ONE

THE AIAS

AI may be used to complete any elements of the task, with students directing AI to achieve the goal. Assessments at this level may also require engagement with AI to achieve goals and so

The AI Assessment Scale (AIAS) has been a core feature on my blog since its inception in early 2023. Resources related to the AIAS are so popular that they account for a sizeable chunk of all web traffic to my site.

Although the AIAS started life as a blog post, it has become much more than that. Through the efforts of coauthors Mike Perkins, Jasper Roe, and Jason MacVaugh the AIAS has grown to be one of the world's most widely used frameworks for assessment design.

All of the academic resources for the AIAS can now be found at the new website: aiassessmentscale.com

This section of this eBook contains all of the articles from my blog. To get the full picture, you should definitely check out aiassessmentscale.com and read the academic publications too.



Perkins, Furze, Roe & MacVaugh (2024). The AI Assessment Scale



The AI Assessment Scale: From no AI to full AI

Published: 2023-04-29 | Original: <https://leonfurze.com/2023/04/29/the-ai-assessment-scale-from-no-ai-to-full-ai/>

AI chatbots like OpenAI's ChatGPT have kicked the hornet's nest in education. From articles declaring the "end of everything" (writing, the English classroom, the entire education system...) to claims that AI will revolutionise education, we've seen the full spectrum of hype.

What we're yet to see much of, however, are practical solutions for dealing with these new technologies. I'm a PhD student, studying AI in education. I'm also a former secondary school teacher, Director of Learning and Teaching, and Head of English. As part of my consulting, I go into schools every week and hear about the challenges of AI and I write *a lot* about how AI is being adopted in schools. I turned one of my blog posts, [Practical Strategies for ChatGPT in Education](#) into a PD which has now run for over 800 educators in secondary and tertiary, online and face to face.

One area where AI is causing a significant stir is assessment. Amidst all of the end-of-everything narrative and fears about students becoming compulsive plagiarists, some schools and universities are trying to adopt sensible approaches to incorporating AI into their assessment tasks.

I've been paying close attention to how people are using AI like ChatGPT. This blog post introduces the AI Assessment Scale, a five-point scale that can help educators determine the level of AI involvement in their assessments.

The AI Assessment Scale: A Closer Look

The AI Assessment Scale ranges from "no AI" to "full AI" and encompasses different levels of AI integration. Here's a breakdown of the five-point scale:

1. **No AI:** The assessment is completed under supervision, and/or handwritten, and/or under exam conditions. This level is suitable for testing knowledge and

comprehension. For example, a traditional multiple-choice exam or an in-class essay written without the use of AI tools.

2. **Brainstorming and ideas:** AI can be used in the initial stages of the assessment for brainstorming and idea generation, checking ideas, etc. This level is suitable for assessments where students need to demonstrate their writing skills, such as constructing their own essays. For instance, students might use AI tools to help generate ideas for a persuasive essay or research paper.
3. **Outlining and notes:** AI can be used to outline entire responses or convert notes (or audio transcriptions) into organised ideas. This level is suitable for assessments where the focus is on the final quality of the writing, word choice, and expression. For example, students might use AI tools to turn their handwritten notes into a cohesive essay outline or restructure their notes to create a more logical flow of ideas.
4. **Feedback and editing:** AI can be used to provide feedback, self-assessment, or editing and revision. This level is ideal when the assessment focuses on the quality of the ideas and understanding, independent of the quality of language and expression. Students might use AI tools to receive instant feedback on their draft essays or identify areas of improvement in their writing.
5. **Full AI:** AI can be used to generate the entire output. This level is suitable when the outcome of the assessment is judged on the earlier organisation, idea generation, discussion, orals, and other methods. For example, students might use AI to create a comprehensive summary of a group discussion or synthesise research findings into a cohesive report.

Let's take a look at each of those in more detail, with a few examples. A PDF of this information can be downloaded for free at the end of this article.

Scale Level	Description	Examples of Assessment Tasks
1. No AI	The assessment is completed under supervision, and/or handwritten, and/	<ol style="list-style-type: none"> 1. Students complete a traditional multiple-choice exam on historical events. 2. Students write an in-class essay about the

	or under exam conditions.	<p>impact of technology on society without the use of AI tools.</p> <p>3. Students solve a series of maths problems on paper during a timed examination.</p>
2. Brainstorming and ideas	AI can be used in the initial stages of the assessment for brainstorming and idea generation.	<ol style="list-style-type: none"> 1. Students use AI to generate ideas for a persuasive essay on the advantages and disadvantages of social media. 2. Students use AI tools to brainstorm potential solutions to an environmental problem in a group project. 3. Students collaborate with AI to develop innovative business ideas for a mock start-up pitch competition.
3. Outlining and notes	AI can be used to outline entire responses or convert notes into organised ideas.	<ol style="list-style-type: none"> 1. Students use AI tools to create an essay outline on the factors contributing to climate change based on their research notes. 2. Students use AI to convert their handwritten notes on a novel into a structured analytical essay outline. 3. Students use AI to organise their research

		findings on public health policies into a clear presentation outline.
4. Feedback and editing	AI can be used to provide feedback, self-assessment, or editing and revision.	<ol style="list-style-type: none"> 1. Students submit their draft essays on the ethical implications of genetic engineering to AI for feedback on structure, clarity, and persuasiveness. 2. Students use AI tools to receive instant feedback on their oral presentations and improve their delivery. 3. Students collaborate with AI to revise and edit their group research papers on the effects of globalisation on local economies.
5. Full AI	AI can be used to generate the entire output.	<ol style="list-style-type: none"> 1. Students provide AI with their research and ideas, then use the AI-generated synthesis to create a comprehensive report on the future of renewable energy. 2. Students input their group discussion notes on the challenges of urban planning into AI to generate a cohesive summary. 3. Students supply AI with their design

		concepts and requirements to generate a visual representation of a proposed architectural project.
--	--	--

Benefits and Challenges of Incorporating AI into Assessments

Incorporating AI into assessments could several potential benefits, such as enhancing the creativity, writing quality, and efficiency of feedback processes and self-assessment. However, there are also challenges to consider. Ensuring that AI does not replace students' critical thinking and problem-solving skills is crucial, and something which I know from experience many parents are concerned about. The [ethical concerns around AI-generated content](#) and balancing the use of AI with traditional assessment methods must be addressed.

Applying the AI Assessment Scale to Non-Writing Tasks

The AI Assessment Scale could easily be applied to other types of assessments, such as oral presentations, group projects, and problem-solving tasks. Furthermore, it can be applied to other AI applications like image generation, data analysis, and virtual simulations. In each case, educators can use the five-point scale to determine the appropriate level of AI integration.

Think about how this Scale could be applied to a visual arts assessment, for example. Students might use no AI in the initial idea generation, and then supplement their brainstorming with an image generation app to look for inspiration – much in the same way students might already use apps like Pinterest. They could create digital artwork themselves, with the use of AI, or using AI to entirely generate the artwork and then editing it themselves.

Creating Clear Guidelines for AI Use in Assessments

The idea behind the AI Assessment Scale isn't to "catch" students in the act of cheating. In fact, [it's getting increasingly difficult to even define what we mean by "cheating"](#). The Scale should be used as a discussion point, and maybe as an addition to a task sheet to clearly indicate how students are permitted or encouraged to use AI for a given task.

It might also be appropriate to break a task down into different elements, and apply different levels of the Scale to different parts of the assessment. For example:

- Your initial brainstorming must be done by hand, on paper, using your knowledge and ideas only. (Level 1)
- Your ideas can then be refined and developed using AI and a thinking routine like Generate-Sort-Connect-Elaborate. (Level 2)
- Once you have developed your initial ideas, your first draft outline should be completed by hand in class. (Level 1)
- You may use AI to expand on your outline and suggest improvements to clarity, logic, and the overall structure. (Levels 3 and 4)
- Once you have completed your draft writing, you may use AI to proofread and make recommendations. (Level 4)

Using ChatGPT with the Assessment Scale

Finally, here are a few examples of how a student might approach an assessment task using prompts in ChatGPT.

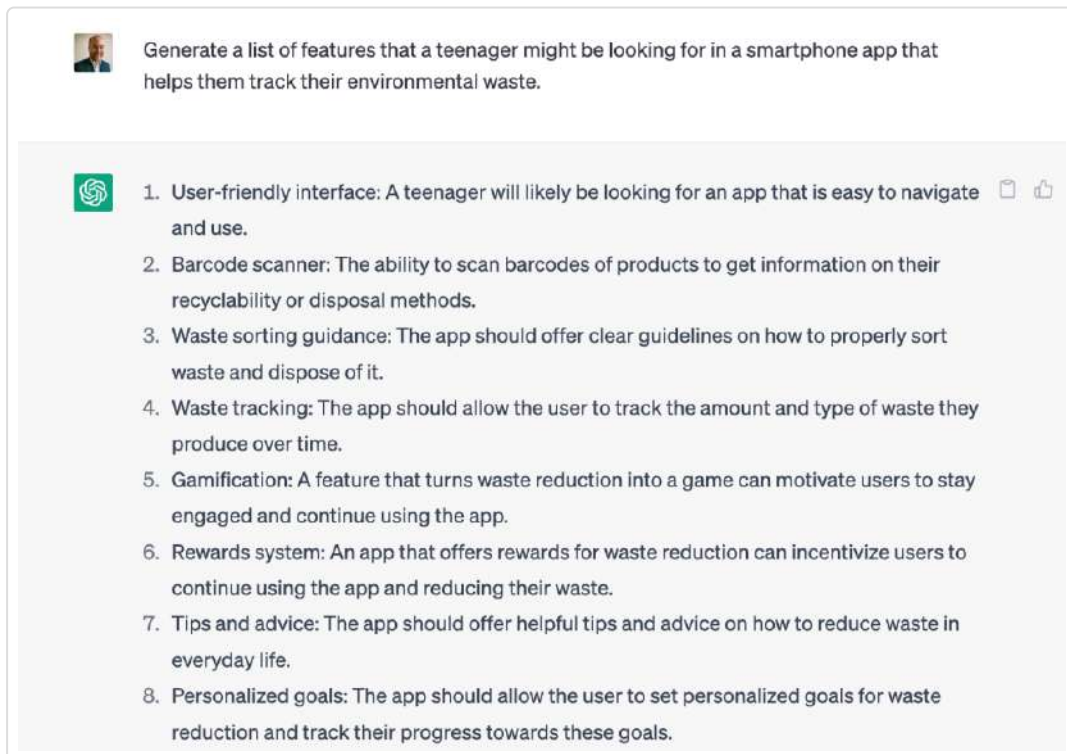
No AI

Sorry, no prompts here!

Brainstorming and ideas

Generate a list of features that a teenager might be looking for in a smartphone app that helps them track their environmental waste.

Here are three of my ideas about <topic>: <copy/paste ideas>. Add onto each of those ideas with three connected ideas each.



The screenshot shows a chat window with a user prompt and an AI response. The user prompt is: "Generate a list of features that a teenager might be looking for in a smartphone app that helps them track their environmental waste." The AI response is a numbered list of 8 features:

1. User-friendly interface: A teenager will likely be looking for an app that is easy to navigate and use.
2. Barcode scanner: The ability to scan barcodes of products to get information on their recyclability or disposal methods.
3. Waste sorting guidance: The app should offer clear guidelines on how to properly sort waste and dispose of it.
4. Waste tracking: The app should allow the user to track the amount and type of waste they produce over time.
5. Gamification: A feature that turns waste reduction into a game can motivate users to stay engaged and continue using the app.
6. Rewards system: An app that offers rewards for waste reduction can incentivize users to continue using the app and reducing their waste.
7. Tips and advice: The app should offer helpful tips and advice on how to reduce waste in everyday life.
8. Personalized goals: The app should allow the user to set personalized goals for waste reduction and track their progress towards these goals.

Outlining and notes

These are my notes from a lesson on <topic>: <copy/paste notes>. Turn them into an outline for an essay on the question <question>

This is a transcript of an audio brain-dump. Turn the transcript into organised notes with headings and subheadings: <copy/paste transcript>

Feedback and editing

Role play: You are my secondary school Year 12 English teacher. You haven't had your morning coffee. Provide some brutally honest feedback about the quality of my arguments: <copy/paste draft>

Check the grammar, spelling, and punctuation of this and make recommendations: <copy/paste draft>



(As your Year 12 English teacher, with a tired and caffeine-deprived brain, I would provide the following feedback):

Alright, let me give you my honest feedback on your arguments. Firstly, your opening statement is a generalization without a clear connection to your topic sentence. You need to provide a smoother transition into your thesis statement.

Regarding your arguments, you make a good point about how AI-powered tutoring programs can help improve student learning outcomes. However, you need to provide more concrete evidence and examples to support this argument. Similarly, your argument about how AI can help reduce the workload of teachers is valid, but it is not clear how AI can provide insights into teaching strategies that are most effective for different types of learners.

Your counterargument is a little weak. While you claim that the concern that AI technologies have the potential to replace human teachers altogether is misguided, you fail to provide a solid rebuttal.

Lastly, you make a valid point about the risks of bias and error in AI technologies, but you need to provide more details and specific examples to support this claim.

Full AI

This is the audio transcript of a group discussion on <topic>. Use it to generate the outline for a PowerPoint, and the outline for a report.

Use the outline for the report you created to write the full report. Begin with the first section and wait for our feedback.



This is the audio transcript of a group discussion. Use it to generate the outline for a PowerPoint, and the outline for a report. [00:00:01] Group member 1: I think AI is definitely unethical and needs to be stopped. The idea of machines becoming more intelligent than humans and potentially taking over the world is just too scary to think about. [00:00:18] Group member 2: But that's just science fiction, isn't it? I mean, we've seen AI technology make incredible advancements in recent years, but that doesn't mean it's going to turn on us and take over the world. We need to be rational about this. [00:00:34] Group member 3: I agree with both of you to some extent. I think there is a real risk of AI becoming unethical if we don't take proper precautions. We need to ensure that AI is programmed with ethical values and that it operates within ethical guidelines. [00:00:51] Group member 1: But who gets to decide what's ethical and what's not? I mean, we can't just leave it up to the programmers and engineers who create these machines. We need to have a broader societal discussion about the ethical implications of AI. [00:01:09] Group member 2: That's a good point. But I also think we need to recognize the potential benefits of AI. It can help us solve complex problems, improve medical diagnoses, and even help us address climate change. We can't just throw out the baby with the bathwater, so to speak. [00:01:27] Group member 3: Absolutely. We need to strike a balance between the potential

The AI Assessment Scale: Version 1

Published: 2023-12-18 | Original: <https://leonfurze.com/2023/12/18/the-ai-assessment-scale-version-1/>

Wait a second! This version of the AI Assessment Scale has now been replaced by a new version. For the most recent 2024 version of the AIAS, please visit the post below

If you're here for information on the development of the first AI Assessment Scale (now renamed version 0 and version 1), then read on!

Back in April I published the [first post on the AI Assessment Scale](#) (AIAS). It emerged from a discussion with the teaching and learning team at Edith Cowan University. We felt that a binary yes/no to using Generative Artificial Intelligence (GenAI) would be unsustainable, and that we needed something with a more realistic approach to integrating the technology into assessments.

A lot has happened since April, and recently I have worked with [Dr Mike Perkins](#), [Dr Jasper Roe](#), and [Dr Jason MacVaugh](#) at British University Vietnam on updating the AIAS accordingly. This post is a reflection on those changes, and the more detailed account we have given in [this article](#).

UPDATE: The AIAS has now been published in a peer reviewed journal, the Journal of University Teaching & Learning Practice (JUTLP). You can find the AIAS article in Vol. 21 Number 6.

Why do we need an AI Assessment Scale?

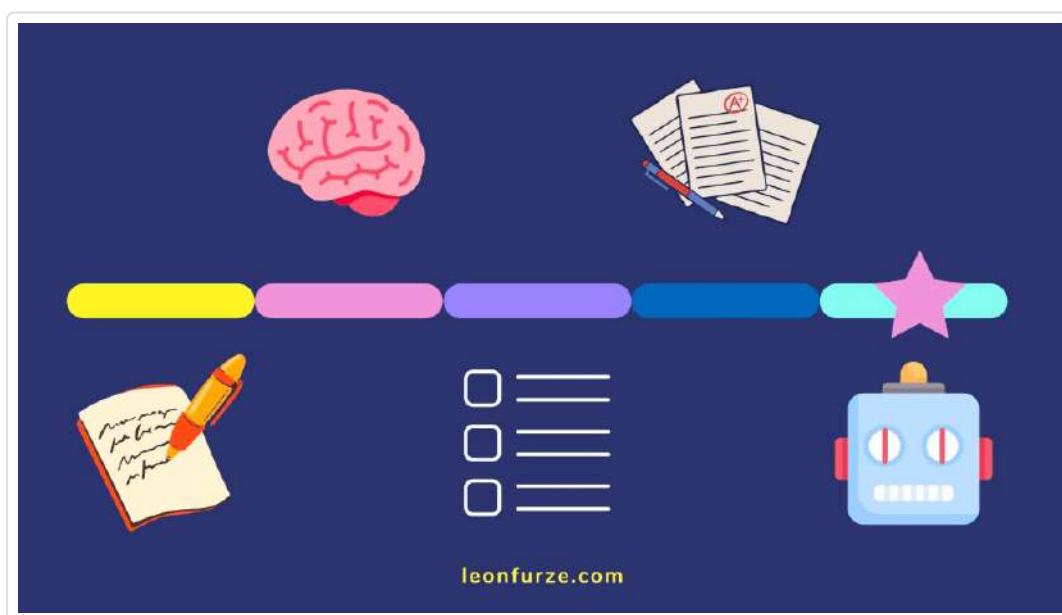
GenAI has caused a stir in education because it has revealed the fragility of many of our current assessments. The narrative of “cheating” that has accompanied the technology since the release of ChatGPT in November 2022 has been overpowering. But we believe that this narrative needs to shift, both to account for the advances in the technology, and to honour our students’ integrity and creativity.

We decide what is and what is not considered cheating. The AIAS is designed to make it easier to articulate those decisions. Also, it is designed to help educators articulate those decisions to students. It is not sufficient to say “the use of ChatGPT is considered cheating”. It’s also inaccurate to label the use of GenAI as plagiarism, as we have explained [in the paper](#).

Detection tools do not work accurately enough for academic misconduct decisions, and so we cannot rely on post-assessment methods of reinforcing academic integrity.

The AIAS, then, comes *before* the assessment task and can be negotiated with students. It allows educators to very clear say *how* and *why* AI can or cannot be used in a given task. You can use the Scale to help students understand that in *this* assessment, it doesn’t matter if you use AI for idea generation or editing, but in *this* assessment we want you to demonstrate your own knowledge and skills unassisted.

The original AIAS



Though the original Scale came from a discussion with tertiary staff, it was grounded in my experience as a secondary English teacher. When I first conceptualised the five levels of the Scale, it was centred on a typical assessment like an essay or analytical response. The five levels were articulated as follows:

1. **No AI:** The assessment is completed under supervision, and/or handwritten, and/or under exam conditions. This level is suitable for testing knowledge and

comprehension. For example, a traditional multiple-choice exam or an in-class essay written without the use of AI tools.

2. **Brainstorming and ideas:** AI can be used in the initial stages of the assessment for brainstorming and idea generation, checking ideas, etc. This level is suitable for assessments where students need to demonstrate their writing skills, such as constructing their own essays. For instance, students might use AI tools to help generate ideas for a persuasive essay or research paper.
3. **Outlining and notes:** AI can be used to outline entire responses or convert notes (or audio transcriptions) into organised ideas. This level is suitable for assessments where the focus is on the final quality of the writing, word choice, and expression. For example, students might use AI tools to turn their handwritten notes into a cohesive essay outline or restructure their notes to create a more logical flow of ideas.
4. **Feedback and editing:** AI can be used to provide feedback, self-assessment, or editing and revision. This level is ideal when the assessment focuses on the quality of the ideas and understanding, independent of the quality of language and expression. Students might use AI tools to receive instant feedback on their draft essays or identify areas of improvement in their writing.
5. **Full AI:** AI can be used to generate the entire output. This level is suitable when the outcome of the assessment is judged on the earlier organisation, idea generation, discussion, orals, and other methods. For example, students might use AI to create a comprehensive summary of a group discussion or synthesise research findings into a cohesive report.

I've taken this AIAS into many K-12 schools that I've worked with this year and have seen it adapted for many contexts beyond the English classroom. I've also discussed it at a tertiary level, which led to conversations with Mike, Jasper and Jason on how it might apply across other disciplines.

Updating the AIAS

Choosing to update the AIAS was really a response to the growing need for a clear and approachable means of addressing GenAI in education, both in K-12 and tertiary. We felt other models, such as the 'traffic light' approach, lacked the nuance needed to clearly articulate to students how AI can and cannot be used in certain tasks. After some discussion of the levels of the AIAS, we produced the following revised Scale:

1	NO AI	<p>The assessment is completed entirely without AI assistance. This level ensures that students rely solely on their knowledge, understanding, and skills.</p> <p style="text-align: center;">AI must not be used at any point during the assessment.</p>
2	AI-ASSISTED IDEA GENERATION AND STRUCTURING	<p>AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.</p> <p style="text-align: center;">No AI content is allowed in the final submission.</p>
3	AI-ASSISTED EDITING	<p>AI can be used to make improvements to the clarity or quality of student created work to improve the final output, but no new content can be created using AI.</p> <p style="text-align: center;">AI can be used, but your original work with no AI content must be provided in an appendix.</p>
4	AI TASK COMPLETION, HUMAN EVALUATION	<p>AI is used to complete certain elements of the task, with students providing discussion or commentary on the AI-generated content. This level requires critical engagement with AI generated content and evaluating its output.</p> <p style="text-align: center;">You will use AI to complete specified tasks in your assessment. Any AI created content must be cited.</p>
5	FULL AI	<p>AI should be used as a 'co-pilot' in order to meet the requirements of the assessment, allowing for a collaborative approach with AI and enhancing creativity.</p> <p style="text-align: center;">You may use AI throughout your assessment to support your own work and do not have to specify which content is AI generated.</p>

The revised AIAS is designed to be flexible enough to account for a variety of disciplines and assessment forms, as well as allowing for the *multimodality* of generative AI. This means it can be applied to the use of [image generation](#) and [editing tools](#) just as easily as [text generation like ChatGPT](#). It can also apply to tasks involving the creation of novel texts (such as writing essays or creating artworks) as well as tasks involving research, summarisation, and synthesis which may (or may not) use [internet-connected GenAI tools](#).

Using the AIAS

In the full paper, we clearly outline each level of the AIAS and the kinds of tasks it may be applied to. Here's a condensed version which explains the five levels of the updated AIAS:

Level 1: No AI

- Students can't use GenAI.
- Suitable for assessments needing personal skills/knowledge.

- Activities: technology-free discussions, in-class work, viva-voce exams.
- Recommended for supervised or low-stakes assessments due to equity concerns.

Level 2: AI-Assisted Idea Generation and Structuring

- GenAI used for brainstorming and structuring ideas, but final work must be human-authored.
- Useful for idea development and foreign language classes.
- Activities: collaborative brainstorming, creating structured outlines, research assistance.

Level 3: AI-Assisted Editing

- Students use GenAI for refining and editing their work.
- Beneficial for language improvements and multimodal content.
- Activities: correcting grammar/spelling, suggesting synonyms, structural edits, visual editing.
- Students submit original work alongside AI-assisted content for authenticity.

Level 4: AI Task Completion, Human Evaluation

- Students actively use GenAI for specific task components, critically evaluate AI outputs.
- Encourages understanding of GenAI's capabilities and limitations.
- Activities: direct AI generation, comparative analysis, critical evaluation, integrating AI content.
- Flexibility in AI and human intelligence interaction.

Level 5: Full AI

- AI used throughout the task at student/teacher discretion.
- Suitable for tasks where GenAI is integral to learning outcomes.
- Activities: co-creation, GenAI exploration, real-time feedback loops, creating GenAI products.
- Encourages exploring GenAI as a collaborative and creative tool.

We also provide supplementary material in the article which includes examples of the AI Assessment Scale applied to a variety of disciplines from Business, to Creative Industries, Computing, and Hospitality. In the supplementary material the five levels are broken down into easy-to-understand tasks with clear advice on how generative AI may or may not be used, and what students need to do in order to reference and acknowledge their use.

Introducing the AI Assessment Scale for Ethical GenAI Assessment

Here is the abstract from the paper, [just released as a preprint](#):

Recent developments in Generative Artificial Intelligence (GenAI) have created a paradigm shift in multiple areas of society, and the use of these technologies is likely to become a defining feature of education in coming decades. GenAI offers transformative pedagogical opportunities, while simultaneously posing ethical and academic challenges. Against this backdrop, we outline a practical, simple, and sufficiently comprehensive tool to allow for the integration of GenAI tools into educational assessment: the AI Assessment Scale (AIAS). The AIAS empowers educators to select the appropriate level of GenAI usage in assessments based on the learning outcomes they seek to address. The AIAS offers greater clarity and transparency for students and educators, provides a fair and equitable policy tool for institutions to work with, and offers a nuanced approach which embraces the opportunities of GenAI while recognising that there are instances where such tools may not be pedagogically appropriate or necessary. By adopting a practical, flexible approach that can be implemented quickly, the AIAS can form a much-needed starting point to address the current uncertainty and anxiety regarding GenAI in education. As a secondary objective, we engage with the current literature and advocate for a refocused discourse on GenAI tools in education, one which foregrounds how technologies can help support and enhance teaching and learning, which contrasts with the current focus on GenAI as a facilitator of academic misconduct.

[ARXIV PREPRINT](#)

As you can see, we have focused on the ethics of generative AI use in a broad sense, as well as the specific concerns of academic conduct and honesty. Our approach is intended to help educators approach generative AI with an open mind and in a way that encourages authentic, secure, and rigorous assessment across disciplines.

Can the AI Assessment Scale stop students "cheating" with AI?

Published: 2024-08-09 | Original: <https://leonfurze.com/2024/08/09/can-the-ai-assessment-scale-stop-students-cheating-with-ai/>

Before you read on: We have updated the AI Assessment Scale and are now focusing our attention on Version 2. However, we understand that many schools and universities around the world have done great things with the original AI Assessment Scale and we continue to support V1. Make sure to check out both versions to decide which is best for you!

Back in April 2023 I published [the first blog post about the 'AI Assessment Scale'](#) which ranged from No AI to Full AI, and was based on my experiences as an English teacher, and a student/educator grappling with the implications of Generative AI.

When Assoc. Prof. Mike Perkins, Dr Jasper Roe, and Assoc. Prof. Jason MacVaugh picked up the scale later in the year and we adapted it to this version, later [published in Journal of University Teaching and Learning Practice \(JUTLP\)](#), the AIAS began to take on a life of its own.

1	NO AI	<p>The assessment is completed entirely without AI assistance. This level ensures that students rely solely on their knowledge, understanding, and skills.</p> <p>AI must not be used at any point during the assessment.</p>
2	AI-ASSISTED IDEA GENERATION AND STRUCTURING	<p>AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.</p> <p>No AI content is allowed in the final submission.</p>
3	AI-ASSISTED EDITING	<p>AI can be used to make improvements to the clarity or quality of student created work to improve the final output, but no new content can be created using AI.</p> <p>AI can be used, but your original work with no AI content must be provided in an appendix.</p>
4	AI TASK COMPLETION, HUMAN EVALUATION	<p>AI is used to complete certain elements of the task, with students providing discussion or commentary on the AI-generated content. This level requires critical engagement with AI generated content and evaluating its output.</p> <p>You will use AI to complete specified tasks in your assessment. Any AI created content must be cited.</p>
5	FULL AI	<p>AI should be used as a 'co-pilot' in order to meet the requirements of the assessment, allowing for a collaborative approach with AI and enhancing creativity.</p> <p>You may use AI throughout your assessment to support your own work and do not have to specify which content is AI generated.</p>

<https://open-publishing.org/journals/index.php/jutlp/article/view/810>

We’ve also seen many adaptations of the AIAS in both K-12 and Higher Education, reflecting the need for flexible, contextual frameworks that help educators and students to understand how AI can be used appropriately in their work.

AI and Academic Integrity

Since the release of ChatGPT in November 2022 the predominant narrative around AI in education has been around cheating and academic integrity. That has led to a proliferation of “detection tools” and other methods of trying to assure assessment security. I’ve written about the concerns extensively on this blog, since they’re such an obvious point of tension for K-12 and Higher Education providers.

- [Generative AI, plagiarism, and “cheating”](#)

- [AI Detection in Education is a Dead End](#)
- [Teaching Writing in the Age of AI: Assessment and “Cheating”](#)
- [Beyond Cheating: Why the ban and block narrative hides the real threats of ChatGPT in education](#)
- [Ditch the Detectors: Six Ways to Rethink Assessment for Generative Artificial Intelligence](#)

In the JUTLP paper on the AI Assessment Scale, we also address the concerns of Generative AI and academic integrity:

Prior to the public release of GenAI tools, the early 2020s had already seen education stakeholders place a renewed focus on academic misconduct and dishonesty, partly because of the COVID-19 pandemic, which led to perceived increases in cheating on behalf of students and teachers (Roe et al., 2023; Walsh et al., 2021). Simultaneously, an ‘arms race’ (Cole & Kiss, 2000; Roe & Perkins, 2022) between technology-enabled academic misconduct and detection software (for example, automated paraphrasing tools) was already in full swing. In this broader context, the focus on academic integrity violations in the era of the GenAI tools can be seen as one node in a network of existing conversations regarding the accelerating pace of digitalisation in HE and the resultant likelihood of what Dawson calls ‘e-cheating’, i.e. cheating that uses or is enabled by technology (Dawson, 2020, p. 4).

The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment
<https://open-publishing.org/journals/index.php/jutlp/article/view/810>

We designed the AIAS partially in response to these concerns: if we can collectively find a way to distinguish between “appropriate” and “inappropriate” AI use, then we can find ways to move beyond the simplistic (and impossible to enforce) use/don’t use or “ban and block” approaches to Generative AI.

One major advantage of the AIAS was that it offered early “permission” for educators to experiment with AI, while still providing some boundaries for academic integrity conversations with students; and it is meant as a conversation starter – not a means of policing or “catching” students in the act of cheating with AI.

The AIAS encourages transparency and honesty regarding AI use, facilitating open conversations between educators and students. This transparent approach has been recognised by the Tertiary Education Quality and Standards Agency (TEQSA) in their advice to universities, *The evolving risk to academic integrity posed by generative artificial intelligence: Options for immediate action*, which includes the AIAS as one of three examples alongside University of Sydney’s “Two Lanes” approach, and UNSW’s adaptation which merges the two lanes with the AIAS.

There are several prominent examples of frameworks that help to make clear to students and staff alike what is appropriate and inappropriate use of AI in learning and assessment tasks. In many instances, a unit/subject coordinator will be best placed to decide on what is appropriate or inappropriate use of AI in a task. The following frameworks will assist with providing this clarity. If there is to be any enforcement of limits on AI use, the limits must be made explicit. Whatever framework each institution implements, what is particularly important is that clear guidance is available for everyone concerned. It is also critical to delineate what is allowable in learning and what is appropriate for assessed tasks where it needs to be clear what work a student has done themselves.

The evolving risk to academic integrity posed by generative artificial intelligence: Options for immediate action (TEQSA, 2024)

But can it stop students from “cheating”?

Despite our hopes that the AIAS can make use of the technology more transparent, we’re also pragmatic about the risks of generative AI for assessment security and “cheating”. As the TEQSA advice states, there is no clear data on how many students are using AI, but “estimates range from approximately 10% to over 60% of cohorts, with an unknown proportion of this use being inappropriate.”

Realistically, the AIAS – or any other framework, [technological solution](#), or approach to academic integrity – cannot stop students from using Generative AI in ways which might be considered dishonest. Our Level 2, for example, suggests students can use AI for initial note taking, ideas, and organisation, but then the final work must be their own: but how do we guarantee that is the case, especially in light of increasingly capable models?

One response to the problem comes from a paper published today by Phillip Dawson, Margaret Bearman, Mollie Dollinger, and David Boud, which takes a different approach to “cheating” entirely.

In *Validity matters more than cheating* the authors argue convincingly that the concept of cheating is an unproductive frame for academic integrity, and we should instead re-centre the concept of “validity” in assessment. Separating the ethical or values-based aspects of cheating – that cheating is wrong or dishonest – from the assurance of learning means we can avoid the “fundamental attribution error” of ascribing cheating to a student’s individual, unethical choice. Instead, we can look for ways in which the system itself might be “wrong” and not just the student: are the methods of assessment such that “all capable students can complete [the task]”? (p. 7)

For Generative AI, this paper offers a useful perspective: the use of AI is not inherently good or bad, wrong or right. It “becomes unacceptable when it threatens validity” (p. 8), but that does not preclude its use in all tasks. Nor does it mean that “No AI” (Level 1 on the AIAS) is necessarily a judgement on whether AI is ethical or not: it is simply an acknowledgement that, in some situations, for some assessments, the technology cannot coexist with the educator’s judgement of what the student can or cannot do unassisted.

In our original paper, we argued that for “No AI” assessments to work, they most likely have to be conducted under supervised, technology-free conditions: not necessarily exams, but certainly on site and in person. This is for two reasons. Firstly, there is no way to guarantee a student with access to a device (phone, laptop, tablet, pair of Meta Ray Bans...) does not have AI assistance. Secondly, if we attempt to use detection tools or other technology solutions, *we run the risk of creating equity issues*, for example between students who have access to more sophisticated, paid AI products (and who will get away with using them) and those who only have limited access to free tools (and will likely get caught).

Dawson, Bearman, Dollinger and Boud have a clear stance on this issue too:

assessments that depend on students not using artificial intelligence but are incapable of preventing students from doing so, are not particularly useful for high-stakes assessment of learning. (p. 8)

No answers, but progress in the right direction

Ultimately, the advice from TEQSA, the paper from Dawson et al., and our own AIAS offer little in the way of “stopping” cheating. They all, however, point to a necessary reframing of academic integrity that could benefit all students.

Cheating behaviours are nothing new, and generative AI may have contributed to the rising number of students finding ways to bypass learning in their university courses and in K-12 education. But we also need to acknowledge that “cheating” is not just a student-issue; it’s also a systemic problem that reflects issues with our assessments.

Hopefully, these technologies won’t be seen simply as a threat to academic integrity and learning, but as a way to shine a light on some of those bigger picture concerns.

Updating the AI Assessment Scale

Published: 2024-08-28 | Original: <https://leonfurze.com/2024/08/28/updating-the-ai-assessment-scale/>

Thanks for visiting! Before you read on, please make sure you visit our new website at <https://aiassessmentscale.com/>. It's the new home for all of the AIAS related materials, publications, and resources. Thanks!

The graphic displays five circular diagrams representing the stages of the AI Assessment Scale (AIAS):

- 1. No AI**: No artificial intelligence is used in the assessment.
- 2. AI Planning**: The assessor uses AI to generate questions, but the student does not use AI.
- 3. AI Collaboration**: The student uses AI to help with the assessment, but the assessor does not use AI.
- 4. Full AI**: Both the assessor and the student use AI.
- 5. AI Exploration**: The student uses AI to explore the assessment, but the assessor does not use AI.

Below the diagrams is a Creative Commons license icon (CC BY-NC-SA) and the text: Perkins, Furze, Roe & MacVaugh (2024). The AI Assessment Scale

AIAS Authors

<p>Mike Perkins Assoc. Prof. Dr. Mike Perkins heads the Centre for Research & Innovation at British University Vietnam. His research focuses on GenAI's impact in higher education, exploring AI text detectors, attitudes to the technology, and the ethical integration of AI in assessments.</p> <p>Contact Mike</p>	<p>Leon Furze Leon Furze is a consultant, best selling author, and PhD candidate at Deakin University Melbourne with over fifteen years' experience in secondary and tertiary education. His PhD is focused on the implications of Generative Artificial Intelligence for teachers of writing.</p> <p>Contact Leon</p>	<p>Jasper Roe Dr Jasper Roe SFHEA is an Assistant Professor in Digital Literacies and Pedagogies at Durham University. His research focuses on educational technology and artificial intelligence.</p> <p>Contact Jasper</p>	<p>Jason MacVaugh As Dean (Higher Education) at British University Vietnam, Jason champions active and problem-based learning to cultivate engaging and inclusive educational environments, emphasising real-world problem-solving skills. His research spans knowledge and innovation management, comparative HRM, and higher education pedagogy.</p> <p>Contact Jason</p>
--	---	---	--

Our new website <https://aiassessmentscale.com/> now has all the resources for the AIAS

It's been over 12 months since the [first blog post](#) about the AI Assessment Scale, and a lot has changed, both with the technology and with our understandings of how it impacts assessments in K-12 and higher education across a range of disciplines. The

AIAS has been adopted by schools and universities worldwide, and will feature in this year's UNESCO digital week. It has also been promoted as a tool for assessment transparency by the Australian Tertiary Education Quality and Standards Agency, as well as adapted and translated around the world.

But we (the authors, Mike Perkins, Jasper Roe, Jason MacVaugh and me) acknowledge that there is still a lot of work to be done in understanding the implications of generative and attentional intelligence for assessment and learning. The original AIAS and its subsequent formal version (published in JUTLP) represents a moment in time where educational institutions across the world were reaching for something to help with the immediate problems of AI, such as the perceived threat to academic integrity.

1	NO AI	<p>The assessment is completed entirely without AI assistance. This level ensures that students rely solely on their knowledge, understanding, and skills.</p> <p>AI must not be used at any point during the assessment.</p>
2	AI-ASSISTED IDEA GENERATION AND STRUCTURING	<p>AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.</p> <p>No AI content is allowed in the final submission.</p>
3	AI-ASSISTED EDITING	<p>AI can be used to make improvements to the clarity or quality of student created work to improve the final output, but no new content can be created using AI.</p> <p>AI can be used, but your original work with no AI content must be provided in an appendix.</p>
4	AI TASK COMPLETION, HUMAN EVALUATION	<p>AI is used to complete certain elements of the task, with students providing discussion or commentary on the AI-generated content. This level requires critical engagement with AI generated content and evaluating its output.</p> <p>You will use AI to complete specified tasks in your assessment. Any AI created content must be cited.</p>
5	FULL AI	<p>AI should be used as a 'co-pilot' in order to meet the requirements of the assessment, allowing for a collaborative approach with AI and enhancing creativity.</p> <p>You may use AI throughout your assessment to support your own work and do not have to specify which content is AI generated.</p>

The AI Assessment Scale as published in JUTLP: <https://open-publishing.org/journals/index.php/jutlp/article/view/810>

Jason Lodge at University of Queensland and TEQSA refers to these as the acute problems of AI, but we recognise the need for robust frameworks that also tackle the chronic problems brought on in some ways by how we approach ideas of assessment and academic integrity in education.

So we have reflected on all of the versions of the AIAS we have seen across the world in K-12 and higher education. We have sought out critique and engaged with diverse perspectives, from school teachers to students, university lecturers, to disability activists, experts in fields including assessment security, cognitive sciences, and pedagogy.

And over the past months, we have refined and invigorated the AI Assessment Scale to bring it up to speed with our current understandings of generative AI and learning.

Timeline of the AIAS

To look at how we have arrived at this updated version, it's useful to consider where the AIAS came from:

- It began as a discussion with teaching and learning staff at Edith Cowan University about the need for more than a binary “use or don't use” solution for AI and assessment during a time of peak AI threat in early 2023.
- The first version, which I will call Version 0, was largely based on my understanding of assessment as a teacher of English and literacy in K-12 and higher education.
- Version 0 was adapted by Mike Perkins, Jason MacVaugh (British University Vietnam), Jasper Roe (James Cook University Singapore) and me to broaden the applicability across other disciplines and make it more relevant to the international higher education context.
- This version, which I will call Version 1, is the most popular and widely used iteration of the AIAS, and was published in JUTLP and a forthcoming peer-reviewed publication based on the pilot study at British University Vietnam. It's also the version that has been most widely adopted around the world.

In between the academic articles on the AI Assessment Scale, the authors have been taking the work in various directions:

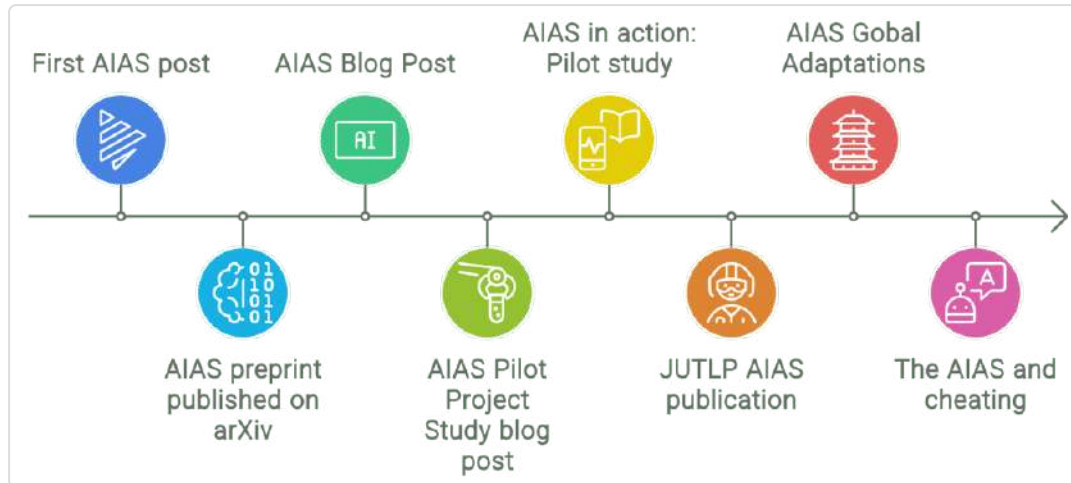
Mike has been presenting the work, seeking more feedback and supporting HE institutions worldwide in implementing the AIAS in various international contexts.

My focus has largely remained on K-12 with occasional forays into higher education, particularly following TEQSA's call for action plans requiring all Australian universities to outline an approach to artificial intelligence. I produced a free ebook containing examples and activities aligned to the levels of the AIAS, and wrote various other blog posts including discussions of the AI Assessment Scale and cheating.

You can access a free ebook on the AIAS with over 50 activities for the 5 levels by signing up for the mailing list here:

Mike and Jasper have continued to refine the AIAS in different contexts, including a recent preprint on how the AIAS can be reframed for English for Academic Purposes (EAP).

- April 2023: [First AIAS post](#)
- December 2023: [AIAS preprint published on arXiv](#)
- December 2023: [AI Assessment Scale Blog Post](#)
- March 2024: [AI Assessment Scale Pilot Project Blog Post](#)
- March 2024: [The AI Assessment Scale Pilot Study preprint](#)
- April 2024: [JUTLP Volume 21 Number 6: AI Assessment Scale Pilot](#)
- May 2024: [AIAS adaptations from across the world](#)
- August 2024: [AIAS for English for Academic Purposes preprint](#)
- August 2024: [The AIAS and “cheating” blog post](#)



Timeline image via Napkin.AI

If you're interested in AI images, audio, and video, make sure to grab the free 20+ page resource [How to Spot a Deepfake](#) by signing up [here](#):

Get the free resource [How to Spot a Deepfake](#)

Refining the AIAS

There are many reasons why we chose to update the AI Assessment Scale. The conversations we have had in the past few months have been long, winding, and complex. In a forthcoming article, we will articulate our main areas for revision in detail, including discussing the theoretical underpinnings of the new levels and our decisions and the discussions we have had with other academics along the way. In this blog post, I'm going to outline some of the main changes from my perspective and why I believe it has been necessary to update the scale.

Technology Changes

Even since the publication of the peer-reviewed version of the [AIAS in JUTLP in April 2024](#) (only four months ago), we have seen some significant advances in the technology which have impacted our thinking.

By the time of publication, GPT-4 was well established. [GPT-4o](#) and [4o Mini](#) have now been released to the public as the base free version of ChatGPT. Similarly, [Claude 3.5 Sonnet](#) is available with limited credits for free users. That the most

powerful models available on the market are now free has had huge ramifications. These tools are significantly more proficient in language, code, mathematical reasoning, and across a range of other benchmarks used to assess the quality of LLMs.

Multimodal generative AI has similarly progressed, and in just the past few weeks, we have seen the release of models like [Flux](#), which can generate images almost undetectable to the human eye, something which I proved recently with a [“real or fake” deep fake game, where out of almost 3000 plays, by the time of writing, fewer than a dozen people have hit 10 out of 10.](#) [Midjourney](#), another incredibly powerful and realistic photo generator announced on August 23 that it was officially moving beyond Discord to its [midjourney.com](#) platform with free credits for all users.

Although OpenAI touted their powerful video generation, Sora, in February, we still have not seen a public release, but models such as [Runway](#) and [Luma](#) are quickly pulling up in terms of quality, and have public versions available. And audio technologies, including the contentious [Udio and Suno](#) and [ElevenLabs](#)’ voice generation models, have established themselves in the past two months.

As educators, we cannot hope to keep up with these technological changes, but we have adapted the AI Assessment Scale to account for them. Most notably, we have removed level four, formerly the “AI + human evaluation” level, and replaced it with the previous level five “Full AI”. At this level, there are no restrictions on how a student chooses to use Artificial Intelligence to approach the task.

The new Level Five is an exploratory level where students, educators, and generative artificial intelligence work together to identify new ways of meeting learning outcomes. As the most recent advice from TEQSA points out, students may well have skills in generative AI that are advanced beyond those of their instructors, and there is no reason not to leverage those skills.

“Engaging in conversations and partnering with students and student groups is critical in developing immediate action. Students bring important perspectives to the discussion about AI as key contributors. In addition, many students are already sophisticated users of these technologies and contribute expert views about how they can and should be used in learning, teaching and assessment.”

The new level five also allows for experimentation with cutting-edge or near-horizon versions of this technology, such as artificial intelligence agents and more advanced multimodal technologies, or the convergence of technologies such as AR/VR and robotics with artificial intelligence.

Pedagogical Changes

We also acknowledge that the ways that people have been teaching with (and sometimes against) AI have developed in the past 18 months in unexpected ways. Universities such as the University of Sydney have invested enormous time and resources in the [creation of open access LLM-based applications such as Cogniti](#), which can be trained on the corpus of an instructor's materials to create a custom chatbot similar to a GPT or a Claude.

Projects leveraging these tools offer incredible opportunities for educators. The original AIAS was seen by many as a tacit permission for educators to use and experiment with artificial intelligence in the classroom, and we have carried that forward into this new version. Again, the only level which rules out artificial intelligence use is Level One: No AI. We now provide more clarity on why educators might choose that level.

To avoid a shopping list of AI tasks, we retain some of the distinction between Levels 2 and 3, articulating Level 2 as assessments where students may use AI as part of the planning, initial composition, research, or in later years of university courses (once they have established foundational skills) through AI-powered research tools. At Level 3, the focus is instead on evaluation, feedback and the use of AI chatbots, such as custom designed course tutors provided by the education institute or the educator. Students can also use AI at Level 3 for the development of their writing, as long as they are still applying a critical eye to the output.

Whilst some approaches like University of Sydney's [two-lane framework](#) are deliberately broad to acknowledge the reality of education and educators' understanding of these technologies, we still believe that people need more support in exploring different applications of generative AI in their contexts.

Assessment Security

One of the most significant changes in our revised AI Assessment Scale is our perspective on assessment security and assessment validity. This has been shaped in large part by discussions in Australia through webinars and materials provided by [TEQSA](#) and [Deakin University's CRADLE](#), and in particular, Phillip Dawson, Margaret Bearman, Mollie Dollinger, and David Boud in their recent publication *Validity matters more than cheating*.

It is a potentially disruptive but absolutely necessary perspective which understands that permitting *any* use of AI effectively permits *all* use of AI, and since it is undetectable and sophisticated across domains, the distinction between previous Levels 2, 3, 4, and even 5 is somewhat arbitrary.

In the revised version, we take the stance that assessment security, in the traditional sense, is only possible at Level 1, but that assessment validity is possible throughout. We will discuss this further in a forthcoming journal article.

Stylistic Changes

A surprising amount of time and attention goes into the branding of academic frameworks for both pragmatic and theoretical reasons: research needs to be communicated clearly, made accessible, and be practical.

The first version of the AIAS was successful in part due to its simplicity and the bold use of the red to green gradient. The AIAS at times has been conflated with traffic light approaches suggesting “No AI” means stop and Green Level Five assessments “Full AI” means go. As the authors have pointed out in numerous publications, including the peer-reviewed JUTLP, the stop-slow-go of the traffic light was never our intention.

Still, the red to green colour scheme of the AI Assessment Scale, whilst being fundamental to its adoption, has troubled us for a while. [In a recent LinkedIn article, Danny Liu of University of Sydney criticised traffic light approaches](#), including ones with “yellow-green, or yellow-red in the mix as well”.

We agree.

So, in the new version, we've adopted more neutral colours, still easily identifiable (e.g., the "pink level", the "blue level"), but without the symbolic connotations of stopping or going. We also chose these colours for accessibility purposes. We rejected the idea of a gradient, since it can be difficult for partially sighted people to discern between levels of the gradient if the contrast between levels is not distinct. We further tested our ideas using online accessibility tools and talking to people with colour blindness and low sight to arrive at the five new distinct colours.

Accessibility and inclusion will be another strand of our forthcoming paper.

Although we have kept the table in the same style as the original AIAS, we also acknowledge that the five levels suggest a hierarchy to some people, perhaps in the sense that "more AI is better", or suggesting that there is a percentage of allowable AI use at a given level (e.g., 25% at level two, 30% at level three and so on).

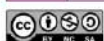
We understand, however, these are arbitrary distinctions. Depending on the context and the discipline and the type of assessment, different levels of AI use may be appropriate at different times. We have therefore designed a new circular version of the Scale graphic, which suggests that all levels may be treated equally in terms of their merits for assessment design.

Mike, Jasper, and Jason will have their own nuanced perspectives and positions on many of these decisions, and we'll draw those together in the full article.

Here is the revised AI Assessment Scale:

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 emphasize 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.



Perkins, Furze, Roe & MacVaugh (2024). The AI Assessment Scale



Our forthcoming paper and the presentations at the UNESCO Digital Week will help to explain our choices further, and provide the theoretical and pedagogical grounding for the revisions. As always, we encourage K-12 and Higher Education educators to both adopt and adapt the AIAS as part of increasing transparency and clarity over the use of AI in learning and assessment.

Thank you to everyone across the world who has contributed to our ideas as we shape and refine the AIAS. We continue to seek feedback from all educators in K-12 and Higher Education.

AIAS: Why We've Driven Through the Traffic Lights

Published: 2024-09-02 | Original: <https://leonfurze.com/2024/09/02/aias-why-weve-driven-through-the-traffic-lights/>

Author's Note: There are four authors for the original and updated AI Assessment Scale (Lead author Assoc. Prof. Mike Perkins of BUV, me, Dr Jasper Roe of JCU Singapore, and Assoc. Prof. Jason MacVaugh of BUV)

In the series of blog posts which will follow, I am expressing my thoughts on the AIAS: though the four of us align on the AIAS I cannot speak for the other authors. For the reflections of all the authors, please read the [published](#) and forthcoming academic papers, or the other authors' own publications and posts.

Last week, [I made a post](#) about our updates to the AI Assessment Scale. These updates are the culmination of using the scale in K-12 and higher education for just over 12 months.

We were also conscious that many schools and universities have already adopted the scale, and we wanted to keep our new version close to the previous so that it did not create a burden for educators choosing to transition from one to the other.

But there is, of course, one immediately obvious change which has already stirred up the most commentary: the move away from the red to green, "traffic light" color scheme.

In this post, I'm going to explain a little more of why we have chosen to replace the traffic lights with what I've been informally calling the "bubble gum colors" (This term might cause my co-authors some amount of horror, but I've said it now so you can't unsee it).

From Traffic Lights to Bubble Gum

“Version zero” of the [AI Assessment Scale](#) had neither clear colours nor the recognisable tabular structure of the published version. In fact, it was a very off-the-cuff affair; a five-point Canva template with a few emojis to indicate tasks at various levels. It was a stop-gap attempt to address the immediate issue of the use/don’t use (or ban and block) approaches to AI in education in early 2023.

But the published and [incredibly popular](#) version, the version that most readers will recognise, used the red to green traffic light colours from the start. Similar traffic light models have emerged over the past 12 months, some adaptations of our AI Assessment Scale, others developing in parallel but not necessarily influenced by our scale.

The connotations of the traffic light colours, however, are clear in every instance of its use. Red means stop, and green means go. Or in the context of artificial intelligence and assessment, red means “do not use”, and green means use freely. In the middle three levels, traffic light systems have an amber level, meaning variously “use with caution,” “use with discretion,” “use critically,” or “some uses but not others are permissible.”

In our original AI assessment scale, we felt the need to break down the yellow level into three distinct categories, allowing for idea generation, editing, and more collaboration with AI systems. We still believe that it is necessary to differentiate between the ways in which educators and students might use artificial intelligence in a given assessment, and I’ll return to that later.

But there are also some unfortunate consequences of the traffic light system that we felt necessary to address.

Breaking the Hierarchy

One problem of the traffic light system is that it visually represents a hierarchy moving from Red through Amber, Yellow, Yellow-Green to Green, and indicating, perhaps to some, a preferential order or taxonomy of skills.

The suggestion in this reading of the AI Assessment Scale is that using AI for idea generation is *better* than no AI, using AI for everything up until the editing is *better*

than just generating ideas, and so on. It also perhaps suggests a taxonomy of the skills outlined in each level, maybe suggesting that idea generation or brainstorming is a lower-order skill than editing and evaluating, or that critical and creative thinking (formally focused on in level four) are more important skills than note-taking and organisation (down in level two).

I'm using a lot of hedging language here, lots of “perhaps” and “maybes” and “suggesting,” because these are all interpretations of the AI assessment scale, and not necessarily our intent as authors. Both as a student and teacher of literature, I understand that the moment you write something and put it out into the world, the reader’s response is as important, if not more, than the author’s intent.

By putting the AIAS out into the world as a traffic light system, we made a design choice. We made a good design choice, since it is striking, memorable, and easy to associate the colours with the different levels, but we also opened the scale up to interpretations of a hierarchy which we don’t believe exists.



Red Means Stop!

The other aspect of the traffic lights we wish to move away from is the suggestion that the AI Assessment Scale can be used to stop or limit the use of AI. We recognised very early that permitting *any* AI use essentially permits *all* AI use, and we never encouraged schools or universities to treat the AI assessment scale as a tool for stopping, catching, or detecting the use of AI.

We tried to make that clear in the [original article](#), where we discussed that the AI Assessment Scale is intended to be used at the beginning of an assessment process – a discussion with students over whether the use of AI is appropriate for a given task. And when we say appropriate, we mean, is it a good use of the technology or a good use of your brain to use AI in this situation?

Again, once you put something out into the world, it can be interpreted in many ways. Some chose to adapt the AI Assessment Scale and try to use it to benchmark the amount of AI use (0% at No AI, 20% at level two, and so on). But the reality is, if you permit students to use AI for brainstorming but ask them to refrain from using it in the final editing of their work, [there is no way to prove that students have stuck to those rules](#). It is essentially nothing more than an honour code (and though they're useful, they of course don't stop all students from cheating).

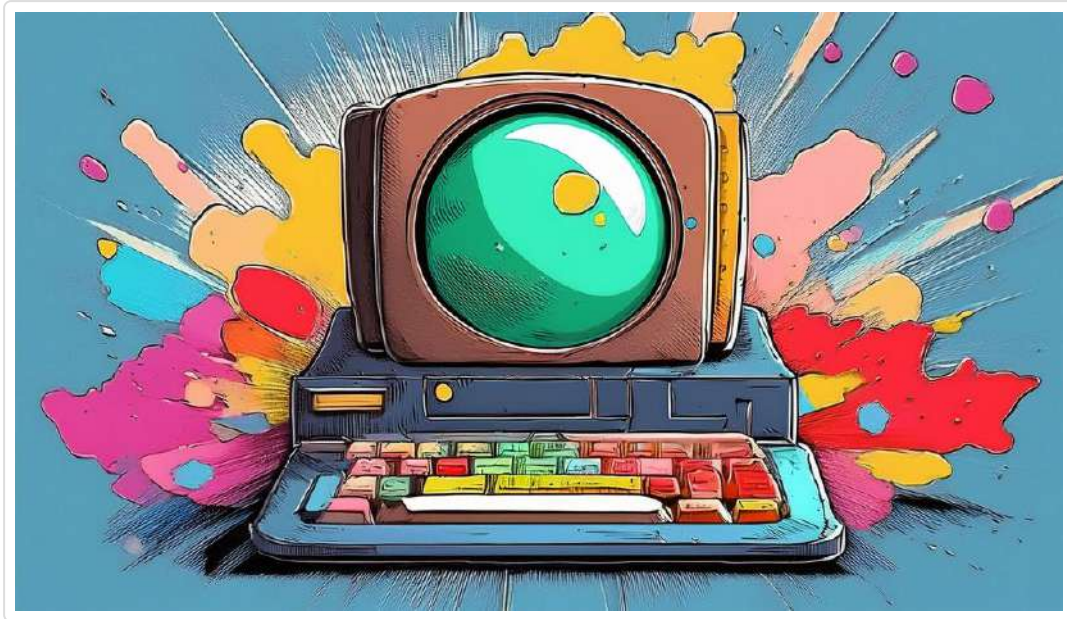
In the new version, we want to make it clear that “No AI” does not equal a glaring red stoplight. “No AI” should be a decision based on what learning needs to be assessed at that moment in time. It should be clearly articulated to students that the reason this given task excludes AI is because it is a bad use of the technology or a bad use of their brain...

For the other levels, just like we are avoiding a taxonomy of skills, we also want to avoid the idea that a student will automatically *stop* using AI at a given point, because we know that assessment processes are complex and iterative. In one of my earlier books, [“Practical Writing Strategies,”](#) co-authored with Benjamin White, we make very clear that the writing process, for example, is a looping, iterative process.

The AI Assessment Scale is not and never was a tool designed to stop or limit the use of AI. It is a scale designed to help educators articulate and teach the appropriate use of AI. Hence, at level two, we design assessment tasks where students are permitted to use AI but have also been explicitly taught *how to use* artificial intelligence as part of the initial design and research stages of their work.

Over time, I will write posts giving extensive examples of what this might look like across disciplines. But for now, here's a brief idea:

Already research tools like [Elicit](#) are available to use artificial intelligence for searching for academic articles based on research questions and synthesising abstracts. These systems are far from perfect, but they offer an interesting alternative or addition to a traditional library search. Whilst traditional library searches require an understanding of complex search terms and Boolean operators (AND, NOT, OR), these AI-powered research tools use semantic understanding and can search based on natural language research questions.



In a first-year university course, you might want students to rely on “old fashioned” research methods and discourage the reliance on AI-powered tools. But by the second or third year of a university course, I would certainly expect students to be experimenting with and learning the strengths and limitations of these AI library tools. In order for students to do that and do it well, they need instruction. And in order to appropriately instruct students on how to use those tools, educators must first be given “permission” to use AI in the classroom.

This is what the AI assessment scale allows for. I’d even go as far to say that level one, which excludes the use of AI, is still a permissive level. Level one gives permission for educators to say, “We don’t have to pour that AI special sauce all over everything. Some things are best taught without AI. Some things are best taught without any technology at all.”

We give the students permission to explore their ideas to the fullest without the distractions of technology. We give permission for long, thoughtful discussions and interactions and engagement, arguments and debate. We give permission for the broadly acknowledged importance of sometimes picking up a pen and writing thoughts by hand as opposed to hammering them out via a keyboard or prompting for them via a chatbot.

I’d encourage anyone using the AI Assessment Scale to consider what each level opens up as opportunities for students to explore technologies or methods of demonstrating their learning.

Hubba Bubba

The last question, then, when we decide to move away from the traffic lights is, what should we replace it with? We had a few options here.

We could, of course, go with a plain text, black and white table, but that doesn't offer much distinction between the levels and as a design choice, it does nothing to make the AI Assessment Scale memorable or visually appealing.

We considered a gradient, but if the contrast between levels is not steep enough, a gradient is less accessible to partially sighted readers. (The gradients between the yellow levels in the first version of the scale, for example, was not steep enough and problematic for some readers who found that levels two, three, and four blended together.)

Taking both the non-hierarchical nature of the updated scale and the visibly distinct levels means we need clearly alternating colors. And moving away from the traffic lights means choosing colors which have a less immediately apparent symbolic connotation.

Enter the bubble gum pastels.

Co-authors Mike Perkins and Jasper Roe, and their colleagues, [have already experimented with a pastel version of the AI Assessment Scale](#) in a recent publication on the AIAS for English for Academic Purposes. They write:





“We recognise that the broad range of tasks present in EAP assessments means that these identified scale levels might be used at different times and for different purposes throughout EAP education, and that no level of the scale is inherently ‘better’ or more suitable than another. We have therefore adjusted the original colour scheme presented in the AIAS from a Red-Green scale to a more neutral palette of colours.”

That pastel colour scheme was the first step in the direction away from the traffic lights, and as we later noticed, harkens back to the original colour schemes on emoji AIAS version zero.

Level	Description	Focus	Example Tasks
Level 1: No AI Use	All language and skills tasks completed without AI assistance.	Developing core language skills and academic competencies independently.	Traditional examinations, in-class presentations, in-class comprehension and critical thinking tasks.
Level 2: AI-Assisted Language Input	AI used to generate or augment input materials.	Enhancing comprehension and analysis skills.	Inviting learners to engage with or create AI-generated texts for reading or listening comprehension and micro skills development. Instructor-created AI materials for assessment or practical use.
Level 3: AI for Limited Language or skills Practice	AI used for targeted practice of specific language and discourse features or academic skills development.	Reinforcing particular aspects of language, discourse, academic or discipline-specific conventions.	AI-generated content for controlled or semi-controlled practice of discipline-specific vocabulary and/or discourse features; simulated academic discussions with AI.
Level 4: AI-Assisted Task Completion with Critical Evaluation	Students use AI to assist in complex academic tasks but must critically evaluate and substantially revise AI outputs.	Developing critical thinking and digital literacy alongside language and academic skills.	Using AI for initial research or drafting, followed by substantial human revision, reflection, evaluation and critique.
Level 5: Selective AI Integration for Advanced Skills	More extensive AI use allowed but emphasizing its role in enhancing, not replacing, student work.	Preparing students for real-world academic and discipline-specific scenarios involving AI.	Using AI for data analysis in research projects, developing AI-enhanced academic presentations and discipline-specific outcome tasks

Table 1: The EAP-ALIAS

We laboured over the exact colours to the point where I can confidently rattle off the hex codes for each of the levels. We tested the colours against online [accessibility tools in Adobe](#), we used Claude to create an artifact to test the accessibility ([you can try it out for yourself here](#)), and we spoke to partially sighted and colourblind individuals and asked them if they could distinguish between the levels and whether the text was legible against the various colour backgrounds.

 #64DEFF	<h2>AIAS Colours and Fonts</h2> <p>If you're not adapting the AIAS with your own colours, these are our choices. We also use the Futura font in various sizes, with bold font on the circle diagrams, titles, and the student instructions in the table.</p>
 #C1FFD2	
 #C1CFFF	
 #FFF1C1	
 #FFC1EE	

Colours and fonts galore from the Design Assets pack, which we have made [freely available via this link](#).

For consistency, we chose colors where the text did not have to change from black to white, which meant that on the darker colors, we needed to be selective about the brightness and the hue.

We are not any of us graphic designers. So we also enlisted the help of the [Fiverr freelancer](#) I have used in the past to compile my blog posts into free PDF ebooks. Even after that design work, we continued to play around with the colours (“play” is a strong word here for the hours of squinting and painstaking Canva work).

Goodbye, Traffic Lights

Honestly, I think this is the longest post I've ever written about choosing five colours. The *content* of the levels is probably more important from a pedagogical point of view than the design decisions. But we know that the first and most obvious change to the AIAS is the new colour scheme, and we know that people responded well to, but also sometimes misinterpreted, the traffic lights.

If you've read this far through my winding discourse about skill taxonomies, accessibility, and hex codes: Congratulations!

And a final request from me: [The AIAS is used the world over](#). It pops up in surprising places, and sometimes the other authors and I are not involved in the conversations and contexts where it is used. That's fine, but if you see criticisms of the scale based on some of the misinterpretations I've listed in this article, particularly around the hierarchy or suggested taxonomy of skills, point them to this post or, better yet, direct them to reach out to me or any of the authors with questions.

We will continue to work with and adapt the AI assessment scale, and we encourage others to do the same. Soon, we will release multiple versions of the scale for adaptation and transformation under their creative commons license.

Thank you, as always, for all of the conversations, examples, criticisms, and discussions around the AIAS. They have all contributed to our updates and continue to make the AIAS flexible and robust in helping educators deal with the implications of generative AI.

AI Assessment Scale (AIAS)

Translations: 2025 Updates

Published: 2025-03-19 | Original: <https://leonfurze.com/2025/03/19/ai-assessment-scale-aias-translations-2025-updates/>

This is an update of an earlier post on translations of the AI Assessment Scale. Please keep sending us your translations as we keep these resources up to date!

The AI Assessment Scale has been adopted across the world in both the original (traffic lights) version, and the updated (bubblegum) version. We have been amazed by how many K-12 and Higher Education institutions have used the AIAS to help their faculty and students.

Since we released the version 2 updates as a CC BY NC SA open access document, and [made the design assets available via this Canva link](#), we have seen many translations of the AIAS across the world.

This post curates all of those translations. If we have missed any, please make sure to get in touch via the form at the bottom of the page.

Before getting into the translations, here's an up to date list of our publications and resources.





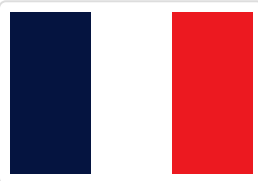


- April 2023: [First AIAS post](#)
- December 2023: [AIAS preprint published on arXiv](#)
- December 2023: [AI Assessment Scale Blog Post](#)
- March 2024: [AI Assessment Scale Pilot Project Blog Post](#)
- March 2024: [The AI Assessment Scale Pilot Study preprint](#)
- April 2024: [JUTLP Volume 21 Number 6: AI Assessment Scale](#)
- May 2024: [AIAS adaptations from across the world blog post](#)
- August 2024: [AIAS for English for Academic Purposes preprint](#)
- August 2024: [The AIAS and “cheating” blog post](#)




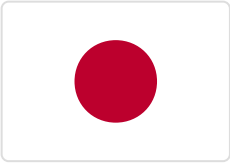


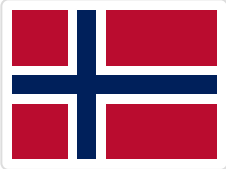





- August 2024: [Updating the AIAS: AIASv2 blog post](#)
- September 2024: [Why We've Driven Through the Traffic Lights blog post](#)
- October 2024: [AJET: Pilot Study Publication blog post](#)
- December 2024: [The AI Assessment Scale Revisited: A Framework for Educational Assessment \(preprint\)](#)

Translating the AIAS

We are indebted to the global community of educators who have translated the AIAS. The following country flags will take you to the translated versions of the AIAS, and below you will find images and examples of the various translations.

Country flag images are taken from Wikipedia Commons and [this github repository](#).

 <p>Brazil – Elton Vinicius Silva</p>	 <p>China (Simplified) – HK Polytechnic: Educational Development Centre</p>	<p>China (Traditional) – HK Polytechnic: Educational Development Centre</p>	 <p>Croatia – Arjana Blažič</p>
 <p>Finland – Kari Kivinen</p>	 <p>France – François Jourde</p>	 <p>Germany – Christine Stoltz</p>	 <p>Greece –</p>

			Anastasia Theodorou
			
Hungary – Éva Tóth and Amália Bognár	Indonesia – Ferdian Satriawan	Italy – Vivaldo Moscatelli	Japan – Yu Urata
			
Netherlands – Thomas Deckers	Netherlands – Sjoerd van Gorp	Norway – Hans Brox	Philippines – Ray Gapuz
			
Poland – Joanna Mytnik	Portugal – Sandra Feliciano	Portugal – Angella Monteiro Santiago	Spain (Catalan) – Miquel Àngel Fuentes

			
<p>Spain (Spanish) – Tatiana Torres</p>	<p>Sweden – Jonatan Tensetti</p>	<p>Turkey – Merve Selen Erkan</p>	<p>Ukraine – Olena Titova</p>
			
<p>United Kingdom – Mike Perkins et al. (English)</p>	<p>Vietnam – Doan Thuy Duong</p>	<p>Wales – Bryony Evelt Hackfort</p>	<p>Romania – Gabriela Grosseck</p>

Where possible, I will provide the link to the editable version of the translated AIAS. Please note that it is always preferable to **make a copy** before editing – some of the translators have created Canva templates, and others have provided the direct link. Please respect the originals if making copies and adjustments, and please respect the CC BY NC SA license.

- [Angella Monteiro Santiago – Portuguese Translation](#)
- [François Jourde – French Translation](#)
- [Tatiana Torres – Spanish Translation](#)
- [Jonatan Tensetti – Swedish Translation](#)
- [Christine Stoltz – German Translation](#)
- [Sjoerd van Gulp – Dutch Translation](#)
- [Thomas Deckers – Dutch Translation](#)
- [Vivaldo Moscatelli – Italian Translation](#)


- Elton Vinicius Silva – Brazilian Portuguese Translation
- Yu Urata – Japanese Translation
- Joanna Mytnik – Polish Translation
- Merve Selen Erkan – Turkish Translation
- Doan Thuy Duong – Vietnamese Translation
- Arjana Blažič – Croatia Translation
- Kari Kivinen – Finnish Translation
- Hans Brox – Norwegian Translation
- Ferdian Satriawan – Indonesian Translation
- Anastasia Theodorou – Greek Translation
- Olena Titova – Ukrainian Translation
- Miquel Àngel Fuentes – Catalan Translation
- Ray Gapuz – Filipino Translation
- Bryony Evett Hackfort – Welsh Translation
- Yumi Kim – Korean Translation
- Éva Tóth and Amália Bognár – Hungarian Translation
- Hong Kong Polytechnic University: Educational Development Centre – Traditional Chinese Translation
- Hong Kong Polytechnic University: Educational Development Centre – Simplified Chinese Translation
- Sandra Feliciano – Portuguese Translation
- Gabriela Grossec – Romanian Translation

How I use the AI Assessment Scale: Part 1

Published: 2025-06-11 | Original: <https://leonfurze.com/2025/06/11/how-i-use-the-ai-assessment-scale-part-1/>

In case you missed it, [we just launched a new website](#) to be the new home for all of our major publications and resources. The AI Assessment Scale has already been used and adapted by hundreds of education providers in K-12, higher education, and adult education worldwide, as well as industry adaptations and [30 translations](#).

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.

 Perkins, Furze, Roe & MacVaugh (2024). The AI Assessment Scale

<https://aiassessmentscale.com>

The AI Assessment Scale is an *idea*, and we believe it is important that ideas are shared freely. That's why, since the first publication, all of our resources regarding the AI Assessment Scale have been published under Open Access Licenses. [You can find, make copies of, and edit all of the current AI Assessment Scale resources at this link to our CC BY-NC-SA 4.0 licensed materials](#), because ideas are open to everyone.

We have seen many great interpretations of the AI Assessment Scale. In this post, I discuss how I would implement the AIAS if I were responsible for rolling it out in a faculty or organisation. I have also checked in with co-authors Mike Perkins and Jasper Roe on this article, since we each have our own interpretations and examples.

This article is not a complete “how-to” guide, but it does articulate a lot of the thinking behind the scale, and provides examples of how I would personally apply it.

The Rationale for the AI Assessment Scale

We have written about this in [each of our major publications](#), but at its core, the AI Assessment Scale developed from conversations with educators about the need for more than a binary “use or don’t use” approach to GenAI.

In 2023 we understood that both educators and students were looking for advice and support for using AI appropriately across various disciplines. So, I wrote the original [AI Assessment Scale](#) blog post with written assessments in mind and based on my experience of 15 years teaching secondary and tertiary English, literature, and literacy.

The rationale to the AI Assessment Scale is this: While we are coming to terms with this technology, both educators and students need support in understanding where and how these technologies might be useful, or where they might be best avoided. One way to provide that support is by breaking assessment task types into common categories familiar to a range of disciplines.

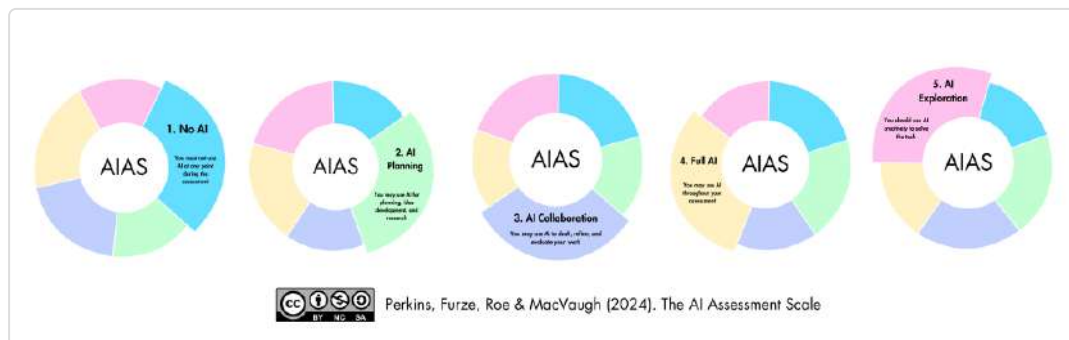
These categories are the levels of our scale.

[In version two of the scale](#), certain other needs became apparent. Among these, the most important was the need for transparency and discourse with students. The AI Assessment Scale aims to help educators and students have transparent discussions about what they believe are appropriate and inappropriate uses of artificial intelligence.

Finally, we see the AI Assessment Scale primarily as an assessment design tool where educators who are experts in their disciplines can review their existing and forthcoming assessments in light of GenAI and make judgements based on their subject knowledge and an understanding of the strengths and limitations of the technology regarding the best use (or non-use) of AI in a given context.

To sum up, my personal “core principles” regarding the AI Assessment Scale are:

1. It reflects the need for a more nuanced understanding beyond a dichotomous “use or don’t use” approach
2. It facilitates transparent communication between teachers and students about what is appropriate and useful and why
3. It is primarily an assessment design tool to be used in the process of discussing, creating, and updating assessments with GenAI in mind



What The AIAS is *not*

The flip side of this discussion is “what the AIAS is *not*“. Again, the AIAS is just an idea, and you are free to interpret it however you like. These are my interpretations of what the AIAS is not:

An assessment security instrument

The AI Assessment Scale is not a tool for [assessment security](#). Beyond our recommendation that Level One (No AI) assessments are conducted technology-free where possible, we do not make any recommendations regarding assessment security. We acknowledge this in [version two](#), where we said that from levels two to five, any use of artificial intelligence essentially permits *all* use of artificial intelligence.

[This is because there is no way to reliably detect AI use](#). Therefore it is impossible to say to a student, “this is a level two task, you may use AI for planning, but you must then promise to stop.” Adequate assessment design choices, especially at level two,

would help to support this, [but nothing can guarantee a student won't use AI beyond your guidelines](#).

In our [original incarnation](#), the AI Assessment Scale had traffic light colours implying stop, slow, go. Our deliberate decision to move away from those colors supports our constantly evolving understanding of the reality of these technologies. I explained those decisions in full in my article, "[Why We've Driven Through the Traffic Lights](#)," and we discuss it in our [version two updates](#).

Personally, I am not a fan of the term "assessment security". My background in K-12 probably has something to do with this, since the term appears to be far more common in Higher Education. It is a term which brings to my mind images of surveillance and adversarial behaviours. However, I have had many conversations with assessment security experts in Australia and they are generally opposed to surveilling technologies, heavy-handed "[catch and smack](#)" methods, and other forms of "[security theatre](#)".

The point is that neither I nor the other authors see the AIAS as a tool for assessment security: you cannot simply show the scale to a student and hope that they will not break your trust.

A shopping list of ideas

The AI Assessment Scale is designed to support educators working across a range of disciplines and areas of education including in K-12, vocational education, and higher education. This means that we try to be supportive of different assessment tasks and types, whilst avoiding being overly prescriptive on uses of GenAI. We also know that GenAI has changed rapidly over the past few years, and any recommendations that we make over particular apps or approaches will quickly become redundant. I personally believe "prompt engineering" and [the use of chatbots as tutors](#) are both short-lived applications of GenAI.

As such, the AI Assessment Scale is not a shopping list of ways to use GenAI. I understand the appeal of providing educators with lists of example prompts or potential use cases such as role playing, critique, spelling, grammar and punctuation, and so on, but I believe that the best way to learn how to use the technology is through experimentation within a discipline.

In the past, I have published resources along those lines myself, including the very popular [Practical Strategies for AI](#) series that later became the basis for [my book](#). Whilst these form part of professional learning for educators, I do not think they are helpful in a framework like the AI Assessment Scale, since there are far too many and varied potential uses of multimodal GenAI across different disciplines, and we are not experts in every discipline.

Instead, the AI Assessment Scale suggests that the disciplinary experts – the educators working alongside the students – should be ultimately responsible for determining how GenAI is or is not used in the context of their subjects’ assessments. However, this must be done with sufficient professional learning and support to understand the strengths and limitations of GenAI so that staff are not working based on opinions or conjecture about what the technology can or can’t do.

There must be a thoughtful balancing of domain expertise and technological expertise. We cannot simply throw a list of prompts at educators and call the job done.

A benchmark

In a similar fashion to the comment on assessment security, I do not see the AI Assessment Scale as a benchmark for students, for example, “at level 2, 20% of your assessment must use AI.” Other than level five, where we say the student should use AI creatively, I would personally recommend against making the use of AI a criteria for the assessment.

Given the contentious nature of the technology and the fact that some students may object to its use on moral or ethical grounds, I never recommend the AIAS is used as an imperative, e.g., “you will be assessed on your use of AI as a research tool,” unless the assessment in question comes from a discipline where artificial intelligence is a required syllabus outcome.

At levels 2-4, I don’t think there is a need to require AI use unless you are specifically teaching or using a particular application, in which case that application must be accessible and equitable.

As far as I’m aware, there aren’t many courses yet that relate to the use of large language models and related technologies on such a specific level. In workshops on the AIAS, I always encourage educators to remember to teach what you mean to

teach, and assess what you have taught. If your course is not about teaching how to use ChatGPT, then your assessment should not have a criterion which judges students on their ability to write a prompt.

However, that does not mean the technology shouldn't be explicitly taught. If you are redesigning an assessment at Level 2, you should be prepared to explore suitable AI applications with students in the same way that you would usually teach the recommended software, methods and approaches of any course. Again, this is where the need to balance technological and domain expertise comes into play.

So to sum up these thoughts, there are three things which I believe the AI Assessment Scale is *not*:

1. It is not an assessment security tool and will not stop students from “cheating” or inappropriately using AI
2. It is not a shopping list of prompts or methods to use AI
3. It is not a benchmark for “how much AI to use” or a criterion unless it is a necessary part of the task

How I Use the AI Assessment Scale: Step-by-step

In a future post, I will move on to how I use the scale in practice when I work with faculties in K-12 and Higher Education, and when I work with students. Again, there is some variance between myself, Mike and Jasper, due to the nature of our different teaching experiences. That is fine: the AIAS is designed with exactly that kind of flexibility in mind.

I will explain how I would use the AI Assessment Scale at an organisational level, a faculty level, and an individual teacher level.

Here's a taster of some of the key ideas in the upcoming post:

- Assessments should be broken down into multiple tasks, assessed formally and informally over time
- Some assessments should permit or explore the use of AI, but only when it does not get in the way of the learning outcomes

- You must to be clear on what the learning outcomes are before making the judgement of whether AI gets in the way
- Every assessment cannot be at “Level 1: No AI”. This is unfair to students, unrealistic, and creates an administrative overhead for educators.

Between now and then, I’d encourage you to look around and see how others are using the scale. For instance, Chevalier College in NSW recently shared an example of how they’re bringing together University of Sydney’s Two Lanes approach and the AIAS:

Joanne Kirby · 1st
Assistant Principal @ Chevalier College | Educational Leadership
1w · 🌐

So great to be a participant in our **Chevalier College** PLM this afternoon, expertly led by **Ryan Noonan**, **Vincent Iacono**, and Matt Heard. Our teaching teams are demonstrating flexibility and adaptability in this age of **#AI** in education. We celebrated 'Pockets of Practice' across the school whereby Teachers harnessed the AIAS developed by Mike Perkins, **Leon Furze**, Jasper Roe, and Jason MacVaugh to rethink, and redesign assessments.

2 Lanes of assessment

Lane 1:
Formative or summative assessment. Sit down assessments and exams. Performance conversations. No AI.

Lane 2:
Formative or summative assessment, taking advantage of AI. In the process of creation and application of skills and knowledge.

Lane 2 Assessment:
Leon Furze AI Assessment scale (AIAS)

AI Planning - Ideation, planning, research. Show the use of AI to refine or develop these ideas.

AI collaboration – AI assists with specific defined tasks, e.g. drafting, refining, evaluating work. Demonstrate critical thinking to cross reference, evaluate and modify AI content.

Full AI – AI is used extensively to achieve the assessment requirements. Demonstrate prompting ability to direct the AI to meet the goals and apply critical thinking to evaluate and modify AI content to meet requirements.

AI exploration – AI is used creatively to solve problems and complete tasks. It goes beyond the task to co-design to further learning. AI use should be assessed.

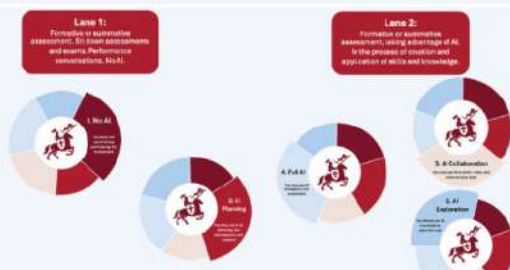

Latency vs No-Latency
In practice, retrieval can occur with varying latency. Use this knowledge to determine the Lane of assessment.

Immediate = No-Latency
Over-time = Latency

No latency = Lane 1
Latency = Lane 2

Lane 1:
Formative or summative assessment. Sit down assessments and exams. Performance conversations. No AI.

Lane 2:
Formative or summative assessment, taking advantage of AI. In the process of creation and application of skills and knowledge.

[Link to images](#)

[Read the post here](#)

PART TWO

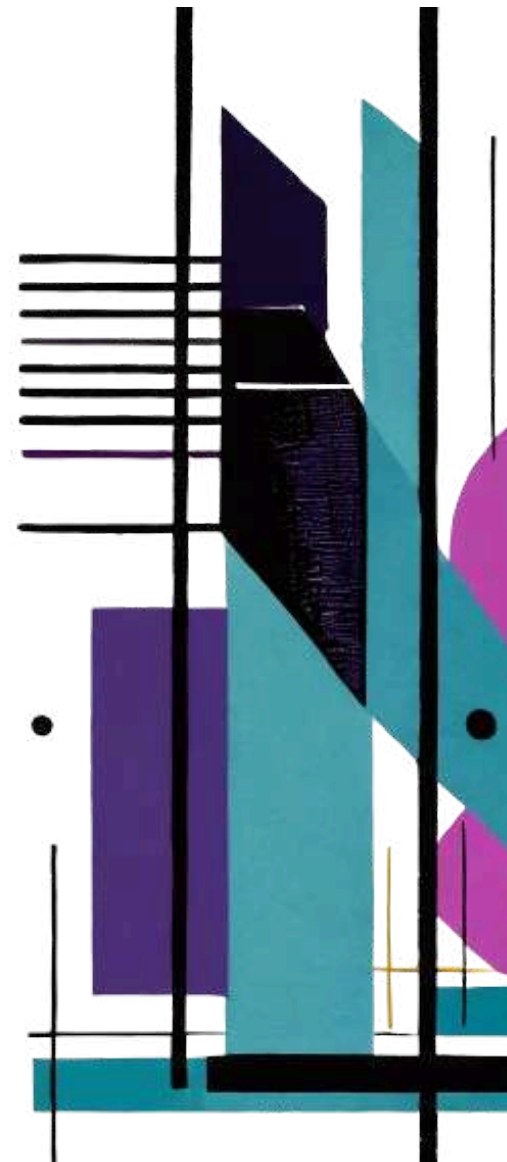
RETHINKING

ASSESSMENT

The second half of this eBook contains articles related to assessment that were published on leonfurze.com/blog between 2023 and 2025. It's a broad range of articles covering:

- AI detection
- Updating methods of assessment for AI
- Ideas for assessing with AI
- Why we should reconsider using AI for grading
- Why take home assessments were a problem long before AI

There are also articles in here which specifically address types of assessment “less vulnerable” to AI. Of course, as the technology changes some of this info might be out of date - stay flexible!



Beyond Cheating: Why the ban and block narrative hides the real threats of ChatGPT in education

Published: 2023-01-21 | Original: <https://leonfurze.com/2023/01/21/beyond-cheating-why-the-ban-and-block-narrative-hides-the-real-threats-of-chatgpt-in-education/>

If you're not familiar with this technology and you'd like to understand the basics about AI, large language models, and ChatGPT, then [check out this post first](#).

This post is part of a series exploring ways we could (and ways we shouldn't) bring large language models like ChatGPT into education. The first post, 'Beyond Efficiency' can be found [here](#):

Amidst all the media hype surrounding ChatGPT, the prevailing fear is that students will use it en masse to cheat in their secondary and tertiary assignments. Aside from the fact that most students haven't actually returned to the classrooms and lecture theatres yet, this narrative obscures more complex threats.

Banning and blocking ChatGPT in education is not just impractical, it's irresponsible. This post explores some of the potential perils of the "cheating machine" narrative in the hope that we can start to explore other options.

Widening the Digital Divide

"Digital divide" is a term that's been thrown around a lot in recent years, particularly as COVID and remote learning highlighted the very real gap between people in our communities who have access to digital technologies, and those who don't. What is less often explored, however, is the impact that policy and approaches to technology in education have on the divide.

One argument of the pro ban/block side is that ChatGPT and technology like it will widen the gap by allowing those with ready access to the technology to gain an unfair advantage over those who don't. But the digital divide is a complex techno-social problem. It's not a simple thing to fix, as failed "[One Laptop Per Child](#)" initiatives and

scorned programs from major tech firms like Meta have proven. Banning ChatGPT won't fix the divide, but it might just make it worse.

In Australia, system-wide technology bans have tended to be implemented by state schools, but not Catholic or Independent sectors. We saw this with [mobile phones](#) in recent years, and with [the blocking of YouTube from Department school networks back in 2007](#). This puts state school students at an immediate disadvantage over students in other sectors, as it denies them access to technologies which may be useful, and which are certainly used in industry and life after school.

More importantly, however, it widens the gap *within* the state sector. Students in state schools who have access to their own devices – whether that's phones, tablets, laptops, or devices at home – will still find ways to use ChatGPT. Just as students flaunted the phone bans and access blocked sites anyway, they'll access the language model despite the ban.

So the only people a ban really impacts are those students who already lack access to devices. Students who, for whatever reason, cannot access a phone or do not have a device at home will be doubly disadvantaged compared to their peers.



The jury is still out over whether banning mobile phones in classrooms is actually effective. Image via Midjourney – prompt in alt text

When Censorship Backfires

We've also seen plenty of instances when attempts to block or censor technologies has the exact opposite effect and makes the subject of the ban more appealing. This isn't just because teenagers are notorious for doing the exact opposite of what they're told: [censorship can lead to an explosion in people seeking alternate ways to access banned technologies](#).

Blocking ChatGPT will only add to the ongoing storm of media hype, quite possible resulting in more students using it to cheat. I've had a few conversations recently with ex-students of mine who are about to start their tertiary studies. By their own admission, they don't read a lot of news. But they've heard of ChatGPT, and almost everything they've heard has revolved around cheating.

By censoring ChatGPT and contributing to the cheating hype, we risk a self-fulfilling prophecy where students are more likely to learn about the technology, seek it out, and use it to cheat on assignments.

Heads in the Sand

There are far greater ethical considerations with AI than academic integrity. The "algorithmic bias" of large language models, for instance, [is well documented](#). Bias inherent in the huge training datasets scraped from the internet makes its way into these models and is then reproduced in the output. This can result in models producing [overtly sexist, racist, ableist, and otherwise discriminatory language](#).

But that's not the only ethical concern. Even seemingly benign output reinforces a static snapshot of the world according to the scraped data. ChatGPT, in its current form, has a dataset that ends in 2021. This means that every output it provides is predicated on a knowledge base that ends in that year. To understand the problem with this, imagine if you were talking to someone whose knowledge ended in 1930, or 1955, or 1972. Their worldview would be significantly different to yours. We expect societal values and norms to change over time, but with current generation language models the [worldview is encoded and static](#), baked in to the technology and reinforced with every output.

Articles outside of the "cheating" narrative have exposed other serious ethical concerns. [A Time magazine piece on the treatment of low-paid Kenyan workers hired](#)

[to sanitise the data](#) shines a light on one of the darker aspects of the technology. To avoid the aforementioned bias, as well as graphic violent and sexual imagery, organisations like OpenAI rely on manual filtering by people working in terrible conditions.

All of these ethical issues and more contribute to the “shadow side” of AI. This doesn’t mean to say that we should ban the technology, however: quite the opposite. If we assume ChatGPT will be used by our students – and it will – then we have an obligation to have open and frank conversations about these concerns.

We can help our students to come to their own conclusions about how, when, and even if to use the technologies rather than simply sticking our heads in the sand and hoping someone else will deal with the bigger problems.

Processing...

Success! You're on the list.

Whoops! There was an error and we couldn't process your subscription. Please reload the page and try again.



Photo by [Anthony Choren](#) on [Unsplash](#)

The Hidden Cost to Teacher Workload

Whenever a technology is banned, the policing and enforcing of that ban ultimately falls to the teachers. Although it will be state and sector bodies who make the decision to block or not, it will be teachers who are faced with challenging students, contacting parents, confiscating devices, and poring over assignments to check for evidence of AI assistance. This isn't just making sure students aren't using calculators under the table in a tech-free Maths test: it will require constant vigilance and a healthy dose of suspicion.

At a time when teacher workload and the teacher shortage crisis is reported on almost as often as this cheating narrative, it seems absurd to add another arbitrary burden to teachers' work. It also perpetuates the "us versus them" narrative of students and educators, with the latter being seen as authoritarian sources of expertise and truth. These technologies are very new, and we have yet to see their full potential. To assume that we have the right to decide for students whether it is an appropriate and acceptable technology seems arrogant, and reminiscent of outdated models of education.

We should be working with students, not against them. ChatGPT, and what it represents, is not another battle that needs to be fought (and probably lost).



Should the burden of policing technology fall to teachers?

Image via Midjourney – prompt in alt-text

The Real Threats of ChatGPT

Will some students use ChatGPT to cheat on assignments? Absolutely. Is that a reason for banning or blocking the technology? Definitely not. The real threats ChatGPT poses to education go beyond issues of academic integrity. Banning the technology will worsen the digital divide, drive more students to use it to cheat, prevent us from educating students of the real ethical concerns, and drive teacher workload even higher.

We can do better.

I want to see schools, universities, and education sectors grappling with the implications of these technologies – the good and the bad. I want to see robust conversations happening in classrooms about the full spectrum of ethical, social, and academic concerns with using AI. I want to see teachers coming up with inventive and creative ways to use the technology to demonstrate to students its capacity beyond writing sub-standard but "passable" essays.

I'd love to see students and teachers working together to create the next generation of these technologies, and helping to shift the attention away from cheating, and towards the things that really matter.

Using ChatGPT for Conferencing and Feedback

Published: 2023-02-08 | Original: <https://leonfurze.com/2023/02/08/using-chatgpt-for-conferencing-and-feedback/>

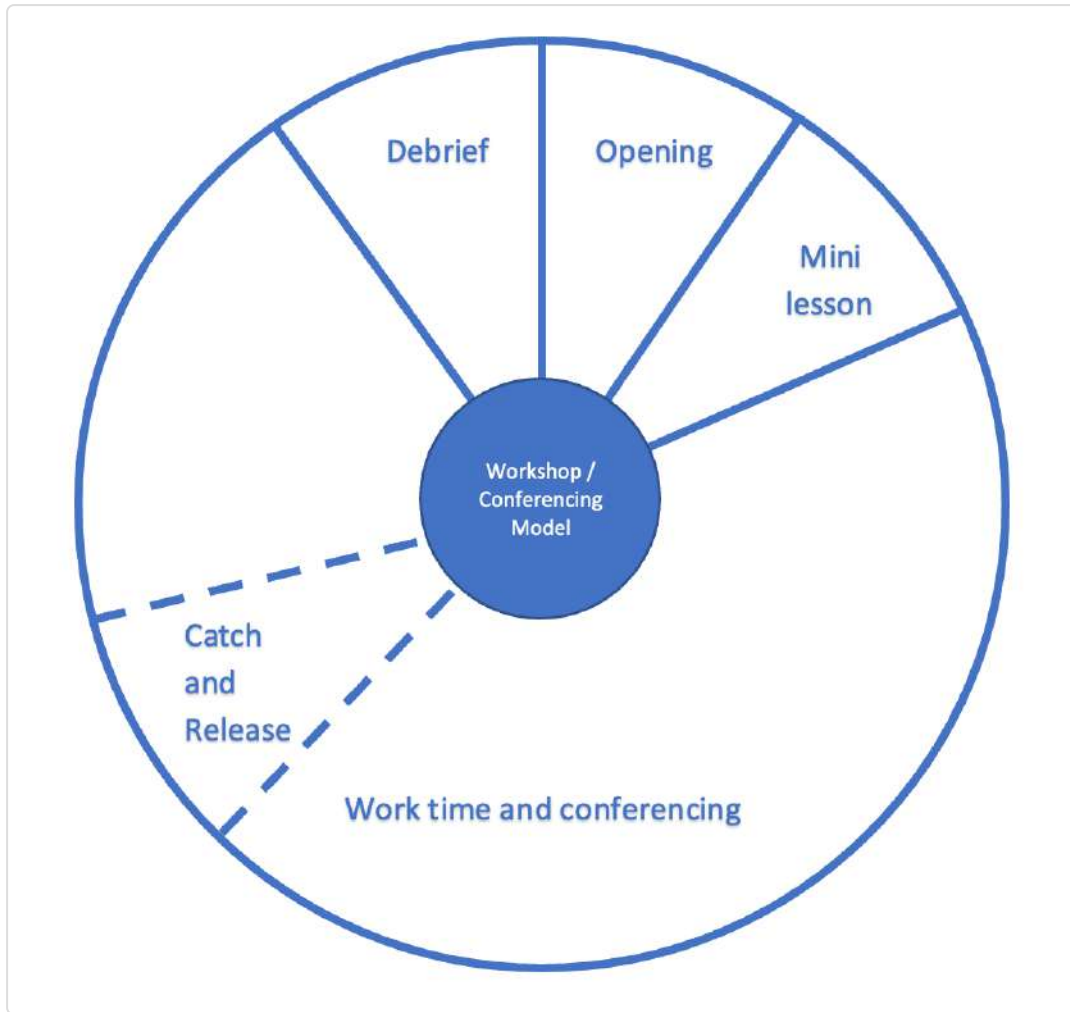
I've used conferencing for years as my main form of feedback and assessment. I stopped collecting piles of books, stopped writing margin notes that no-one ever read, and stopped correcting work like a human spell-checker. Aside from the hours of time saved by not "correcting" work, I also built stronger relationships with students as a result of regularly sitting with them 1:1 to go through their work.

At the moment, ChatGPT has been banned by the Department of Education in most states across Australia. Hopefully there will come a time when this stance is reviewed and where the technology can be used by teachers and students in responsible ways.

One of the ways ChatGPT could be highly effective is as a support technology for conferencing.

What is conferencing?

I use a conferencing model from Cris Tovani, outlined in her book *So, What Do they Really Know?*. [It's also explained in brief here.](#)



The workshop model – Cris Tovani

The workshop/conferencing model of lesson is perfectly suited for providing live feedback. After students have submitted a piece of work – whether it’s a sentence, a paragraph, or a whole essay – I’ll commit to the next one to four lessons using this structure. Here’s an example from Year 12 English:

1. **Opening/Mini lesson:** Overview of ‘tutorials’ available for comparison texts. These tutes were created in the 2020 lockdowns and include an essay and discussion from each of the Y12 teachers on an idea from the text. These resources could also be sourced from study guides, YouTube, or anywhere else suitable. As long as it is work that students can complete under their own steam, it’s fine.
2. **Work time:** Students, individually or in groups, watch the tutorial videos, take their own notes, annotate the sample essays, read the books.

3. **Conferencing:** During the work time, I'll sit at the front of the room or move from student to student, depending on how long I'll need to speak to them for. If it's feedback on a couple of sentences – like a contention, or an outline – I'll go from desk to desk. If it's feedback on a full practice essay, I'll call them up one at a time. I provide my feedback verbally. This is generally the first time I'm reading their work – remember, this is about effectively using my time as well as providing good feedback, so I'm not spending hours out of class hunched over writing feedback and then just reading it to them. It's live, which has the added benefit that my feedback must be concise, and to focus on only one or two key areas for improvement. The students make their own notes on what I'm saying. I never touch the red pen.
4. **Catch and release:** Occasionally I'll pop my head up to see how the class is going, ask if there are any questions about the task at hand, and so on.
5. **Debrief:** A quick, 5 minute discussion at the end of the lesson to check in on the main points of the tutes.

Like I said, for a longer piece of work I'll commit to up to four lessons in this structure. I have twenty-something students in the class, so if I sit with each for ten minutes, there's my 4 x 50-minute periods. That's OK. I'm comfortable giving up four periods of everyone's time to go through this process, and after a few rounds, so are the students.

This conferencing process happens several times per unit. As often as possible really, as I believe the more time spent working one-on-one with students the better, and this is the best way to do that in the time constraints. When the final piece of work is submitted it goes straight to comment/grade feedback, with no corrective annotation. This summative feedback does not serve the purpose of improving the next piece of work, which is generally on a totally different topic.

The goal throughout this process is on constantly developing student autonomy. It achieves this in a few ways. Firstly, the students know not to rely on me as a crutch during the conferencing lessons. If you don't know an answer, Google it (or ask ChatGPT if you're not too worried about accuracy). If you're stuck, ask a friend.

Next, getting the students to write down the feedback, and to reflect on it themselves and develop their own goals for next time, is much more effective than corrective written feedback. Finally, I've noticed that students are much more open to asking for

specific help and advice after a few rounds of this process – it helps build trust, and trust is a prerequisite for asking for, receiving, and responding to feedback.

Processing...

Success! You're on the list.

Whoops! There was an error and we couldn't process your subscription. Please reload the page and try again.

Bringing ChatGPT into Conferencing

If ChatGPT isn't unblocked, it's highly likely that another form of generative AI makes its way into your classroom soon. Whatever the application or platform you use, there are many great ways to incorporate a language model into a conferencing session:

1. **Pre-class Preparation:** Students can use ChatGPT to research and gather information on a topic before the conferencing lesson begins. With the usual caveat of checking for accuracy, ChatGPT can be a great research tool.
2. **Writing Practice:** Students can practice writing and receive immediate feedback from ChatGPT on their writing, including grammar, structure, and clarity. They can get technical feedback instantly while the teacher works on more nuanced feedback.
3. **Group Discussions:** ChatGPT can facilitate group discussions by providing prompts, asking questions, and keeping track of contributions from each member. Small groups of students could work together using ChatGPT as a tutor while the teacher sits with students 1:1.
4. **Conferencing Support:** ChatGPT can assist during the conferencing process by providing additional information and context to support the teacher's feedback. Once the teacher has left, a student can run follow-up questions through ChatGPT.
5. **Personalised Feedback:** ChatGPT can provide individualised feedback to each student during the conferencing process, taking into account their unique needs. For instance if one student is writing a blog and another a short story (e.g., for VCE English Creating Texts) ChatGPT could provide different advice to each.

6. **Error Correction:** ChatGPT can help students identify and correct errors in their writing, giving them the opportunity to make immediate improvements. Students don't need to "wait their turn" for 1:1 feedback over smaller errors.
7. **Vocabulary Expansion:** ChatGPT can suggest new words and phrases to students during the conferencing process, helping them to expand their vocabulary and improve their writing. It's basically an extension of a thesaurus, but it can provide many more examples of how to use a word in context.
8. **Peer Feedback:** ChatGPT can facilitate peer feedback sessions, where students can review and provide feedback to each other's work.
9. **Reflection and Self-Assessment:** ChatGPT can support students' reflection and self-assessment by guiding them through the process and helping them set goals for future improvement.
10. **Progress Tracking:** ChatGPT can keep track of students' progress during the conferencing process, providing them with a clear record of their achievements and areas for improvement. Using the applications capacity to store chats, students could refer back to their ongoing "feedback chat" throughout their studies.

Conferencing is a highly effective method of providing live feedback to students. By using the conferencing model, teachers can build stronger relationships with students, provide concise and focused feedback, and develop student autonomy.

The addition of ChatGPT or other AI-powered language models can further enhance this process, by helping students gather information, practice writing, and receive immediate feedback.

While ChatGPT is currently banned by the Department of Education in many states in Australia, there may come a time in the future when the technology can be used in responsible ways. In the meantime, it's important for teachers to continue to use conferencing and other effective methods of feedback and assessment to support student learning and growth.

Teaching Writing in the Age of AI: Assessment and "Cheating"

Published: 2023-02-18 | Original: <https://leonfurze.com/2023/02/18/teaching-writing-in-the-age-of-ai-assessment-and-cheating/>

In this post, I'll be exploring the assessment of writing, and why AI is such an apparent threat to the way we currently teach and assess. In putting this article together, I leaned on a couple of resources which are well worth checking out:

- [John Warner's article from December 'Freaking out about ChatGPT'](#) presents a level-headed response to writing assessment
- [Phil Dawson's AARE article on "cheating"](#) provides an excellent discussion of why students cheat, and how to address cheating.

[There's quite a bit of discussion here about what constitutes cheating, and ways to build "anti-cheating" approaches into assessment. If you're more interested in the applications for teaching writing, then jump straight to the practical part by clicking here.](#)

Assessment and "cheating"

Concepts of assessment and cheating seem inextricably linked. The media has been awash with concerns about students using models like ChatGPT to cheat in online exams, bypass traditional assessment methods, and write entire essays.

However, the cheating narrative seems to be largely a media invention that hasn't taken into account student or educator voice. While some have expressed concerns about the impact of AI on assessment, it's still not clear whether fears of cheating are reflected by how students actually use the technology.

As for the teachers, many educators have shown a keen interest in exploring the opportunities of AI, as well as its challenges, through professional development and webinars. So where is the fear and speculation about cheating coming from? It

appears to be a convenient narrative for headlines, rather than a reflection of the views of educators and students.

To shift the narrative away from concerns of cheating, it's important to first understand why students might cheat in the first place, and what “cheating” actually means.



“Cheating” has evolved, but we need to focus on why students cheat, not just how.

Photo by RODNAE Productions on [Pexels.com](https://www.pexels.com)

Why students cheat

Research into cheating tends to focus on higher education. At a tertiary level cheating is big business, with “contract cheating” identified as one of the major issues. The Tertiary Education Quality and Standards Agency, TEQSA, defines contract cheating as follows:

...when students outsource their assessments to a third party, whether that is a commercial provider, current or former student, family member or acquaintance. It includes the unauthorised use of file-sharing sites, as well as organising another person to take an examination.

It's such a problem that the [Australian Government](#) has passed laws to penalise contract cheating agencies which provide free or paid services such as producing essays and sitting examinations.

ChatGPT and other AI models, however, might do an even better job than the legal system of putting contract cheating agencies out of business. Why pay up to hundreds of dollars for someone to write an essay on your behalf when GPT can do it for free? The release of ChatGPT has forced us to reconsider traditional methods of assessing student writing, and to question whether relying solely on a single artefact, such as an essay or exam, is the most effective way to evaluate students' learning.

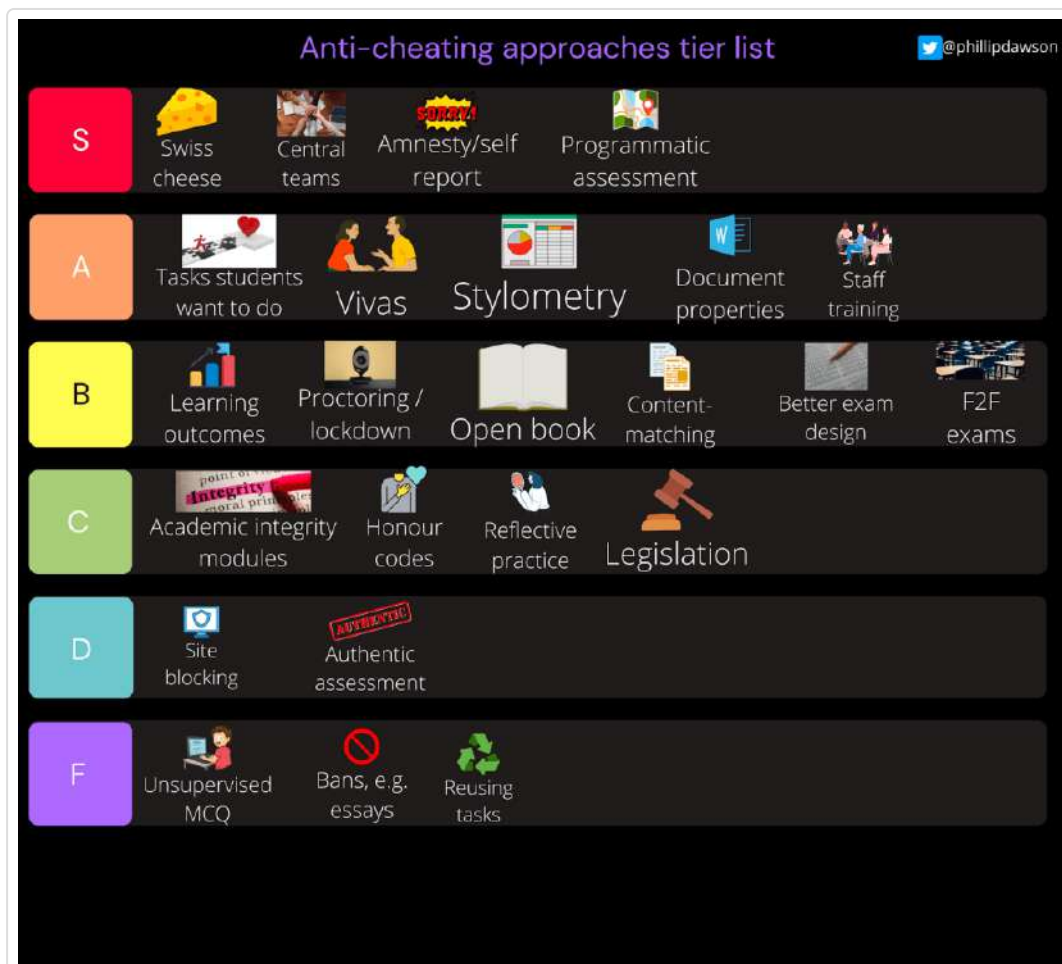
Philip Dawson, Associate Director of the Centre for Research in Assessment and Digital Learning (CRADLE) at Deakin University is more interested in *why* students cheat than finding heavy-handed ways to stop cheating. In his [article for AARE](#), Dawson suggests that there are many reasons why students cheat, including:

- Inadequate preparation or time management skills
- Pressure to achieve high grades or compete with peers
- Lack of engagement with the subject matter or task
- Perception that cheating is easy or that the risks of being caught are low
- Belief that cheating is a victimless crime or that everyone else is doing it
- Lack of understanding or respect for academic integrity
- Availability of technology and resources that make cheating easier.

Dawson notes that these reasons are not mutually exclusive and can interact with each other in complex ways. Although this research is centred on higher education, the conversation is equally important for secondary schools. In fact, several of Dawson's points including engagement, understanding of academic integrity, and pressure to achieve high grades should be things we address in spite of ChatGPT.

The role of assessment design in addressing cheating

Sticking with Dawson’s article for a while, he suggests that there is a “tier list” of approaches to anti-cheating, ranking methods from most down to least effective:



Philip Dawson’s Tier List for Anti-Cheating Approaches

Here’s a further explanation of Dawson’s tiers:

Tiers	Approaches
S (why “S?”)	Swiss cheese (layering multiple anti-cheating interventions), central teams (resourcing cheating experts), amnesty/self-report (students coming forward to confess),

Tiers	Approaches
	programmatic assessment (building best practice into all elements of program design, not just assessment)
A	Tasks students want to do (students less likely to cheat if they enjoy the task), vivas (discussions with students to determine if they did the work themselves), stylometry (tech that compares student assignments to see if they were written by the same person), document properties (signals found in document metadata), staff training (detecting contract cheating with training)
B	Learning outcomes (assessing higher-order outcomes), proctoring/lockdown (using lockdown browsers and other proctoring approaches), open book (exam that allows notes, eliminating one type of cheating), content-matching (text-matching tools to deter copy-paste plagiarism), better exam design (tricks to prevent cheating), face-to-face exams
C	Academic integrity modules (teaching students about integrity), honour codes (students sign document promising to be honest), reflective practice (students reflect on work), legislation (laws banning cheating)
D	Site blocking (banning access to tools like ChatGPT), authentic assessment (application in real-life settings, but with no evidence it reduces cheating)
F	Unsupervised MCQs (multiple choice), bans (banning essays entirely), reusing the same task

Explanation of [Philip Dawson's Tier List for Anti-Cheating Approaches](#)

Unfortunately, much of the discourse around ChatGPT and education has centred on punitive or restrictive methods such as those found in Dawson's D-F tiers. The "Swiss cheese" approach of layering methods – drawing on approaches from across the tiers

– would seem to be a more productive way to deal with the potential misuse of the technology.

These approaches are useful for addressing cheating in higher education, but they could also be considered in the design of assessment tasks in secondary school. By creating tasks that are engaging, relevant, and aligned with learning outcomes, educators can encourage students to approach the task with greater integrity and motivation.

Teaching and Assessing Writing

So how do we combine the topmost, more effective, tiers of Dawson’s approaches with ways of teaching and assessing writing?

In [John Warner’s article](#) he suggested a few ways of preventing students from “cheating” at writing, many of which overlap with Dawson’s ideas:

- Give students learning experiences of intrinsic interest and extrinsic worth so they’re not tempted to cheat.
- Use methods of assessment that take into consideration the processes and experiences of learning, rather than simply relying on a single artefact like an essay or exam.
- Require students to practice metacognitive reflection, asking them to articulate what they have learned and then valuing and responding to what they tell us.
- Change the way we grade and require students to demonstrate the ability to synthesise multiple sources.
- Ask students to bring their own unique perspectives and intelligences to the questions we ask them.
- Create assignments that integrate this technology into the learning.

You can see how tasks which focus on intrinsic motivation align with the A-tier, and considering the entire process of learning overlaps with the “programmatic assessment” of the S-tier. Metacognitive reflections might form part of a viva, or a C-tier reflective response. I’m very excited about the potential for these kinds of approaches in teaching writing.

Here are some practical suggestions for what that might look like in secondary school writing education:

Approach	Example 1	Example 2	Example 3
Intrinsically motivating tasks	Writing a personal essay on a topic of the student's choice	Creating a podcast episode on a current event of interest to the student	Composing a poem or creative piece that captures a particular emotion or experience
Tasks which value the whole process of learning	Keeping a writing journal throughout the semester to track progress and reflect on strengths and weaknesses	Engaging in peer review and revision workshops with a focus on the writing process rather than the final product	Incorporating low-stakes writing assignments that allow for experimentation and exploration of ideas
Tasks which require metacognitive reflection	Writing a reflection essay on the process of writing an assigned essay, including challenges faced and	Recording a video reflecting on the revision process and changes made to an initial draft	Completing a self-evaluation checklist that prompts students to consider their own writing strengths and weaknesses

Approach	Example 1	Example 2	Example 3
	strategies employed		
Assessments which require synthesis of ideas and sources	Composing a research paper that draws upon a variety of sources to develop a nuanced argument	Writing a critical analysis essay that requires students to integrate multiple perspectives on a particular topic	Creating a multimedia presentation that combines text, images, and video to explore a complex issue
Students' own perspectives	Writing an argumentative essay that requires students to take a stance on a current event or controversial topic	Crafting a personal narrative that draws upon the student's unique experiences and worldview	Composing a persuasive speech on an issue that the student is passionate about
Assignments that integrate LLM technology into the learning	Collaborating with a ChatGPT interface to create an outline or generate ideas for an essay	Using GPT-3 language models to create a dialogue between historical figures or	Experimenting with AI-generated writing prompts and using them as a starting point for

Approach	Example 1	Example 2	Example 3
		characters in a novel	creative writing exercises

Avoiding the arms race

We don't need to try to "beat" AI, and it's unreasonable to place the burden on teachers and educators to police the use of these technologies. They will become prolific, replacing or augmenting many of our current generation of technologies including search and word processing.

Engaging in an arms race against AI and falling back on "lower tier" approaches such as blocking, supervised examinations, and legislation may temporarily discourage cheating, but it does not address the reasons why students cheat in the first place.



Multiple Choice Questions (MCQs), blocking site access, and heavy-handed approaches to exams are not the answer.

Photo by Andy Barbour on [Pexels.com](https://www.pexels.com)

To end this article, I'd like to suggest how to layer several approaches in a single unit of work (The S-Tier "Swiss Cheese" approach). These would be particularly useful in

the kind of unit where a student is required to read a text and produce an analytical essay in response (such as VCE English and EAL Unit 2 OC1, or a HSC Band 6 essay). It has been one of the most talked about assessment styles in the cheating discourse, and I want to explore ways we can move beyond that narrative.

Assessment stage	Approaches
Before unit	Academic integrity modules for students, honour codes, staff training in discussing academic integrity and cheating
During unit (formative assessment)	Intrinsically motivating tasks, synthesis of multiple sources including diverse perspectives on the set text, open book assessments
After unit (summative assessment)	“Mini vivas”, reflective practice, use of AI to assist with essay outlines and drafting, opportunities for amnesty/self-report

Here’s what that looks like integrated into a typical 8-week analytical response unit of work:

Week No.	Suggested Lesson activities
Before unit	Establish the importance of academic integrity by introducing students to academic conventions and standards . Teach students how to reference sources, avoid plagiarism, and use citation tools. Discuss “ honour codes ” with students.
Week 1	Introduce the text and provide an overview of the unit. Set clear learning outcomes and assessment criteria for the final tasks. Assign an intrinsically motivating task related

Week No.	Suggested Lesson activities
	to the opening of the text, such as a personal /creative writing exercise or a discussion forum.
Week 2	Continue exploring the text through reading strategies , close reading and annotation. Conduct a formative assessment, such as a group presentation or debate, to encourage collaboration and engagement with the text. Use an open book format for the assessment (e.g., a Socratic seminar with pre-prepared responses). Provide opportunities for metacognitive reflections such as “metacognitive journal” or writing journal.
Week 3	Assign a critical analysis task, such as a close reading or a character analysis. Demonstrate text-matching and AI detection tools (e.g. Turnitin or OpenAI’s detection tool) to used detect plagiarism and ensure originality of student work: discuss the limitations of these tools and how students can use them to detect “accidental plagiarism”. Provide students with feedback and opportunities for revision.
Week 4	Continue close reading activities . Conduct a reflective practice session, where students reflect on their writing so far and its development, and receive feedback from peers and teachers. Encourage students to engage in self-reflection on their understanding of the text.
Week 5	Conduct a face-to-face or pre-recorded “ mini viva “, where students discuss the writing they have produced so far with the teacher and provide evidence of their learning. Use the viva to assess the authenticity of student work and to continue to discuss academic integrity. (Re)Introduce students to the final essay task and discuss analytical

Week No.	Suggested Lesson activities
	essay writing skills, drawing on their prior knowledge of essay tasks.
Week 6	Assign a research task related to the text, where students must synthesise information from multiple sources and demonstrate critical thinking skills. Students should be encouraged or even required to use AI tools such as ChatGPT , perplexity.io , you.com , and new Bing . Discuss the strengths and limitations of these tools compared to tools like traditional search engines, Google Scholar , and library services.
Week 7	Begin the final assessment tasks, such as the analytical essay. Allow class time for students to plan, draft, and generate ideas. Conduct most of the draft writing during class time. Encourage honesty and integrity, and provide support for students who may be struggling with the demands of the unit.
Week 8	Conduct a final reflective practice session, where students reflect on their learning and development over the unit. Provide students with an opportunity to self-report cheating and seek help and support, for example by using a Google Form or other survey. Require students to submit a written reflection on their learning, and provide feedback and opportunities for improvement. Have students submit the final essay.

Secondary School Assessment and Artificial Intelligence

Published: 2023-06-13 | Original: <https://leonfurze.com/2023/06/13/secondary-school-assessment-and-artificial-intelligence/>

When ChatGPT landed on our doorsteps in November 2022, it largely slipped beneath the education radar. By the time term 1 2023 rolled around, however, it's fair to say that the situation had changed. Most states in Australia banned the technology in Department schools, and those bans remained in place for several months.

Despite the bumpy start, policy is quickly catching up with the reality of these technologies. We're now starting to see some guidelines emerging from various states and sectors for the appropriate use of AI in secondary education. In the tertiary sector, we're also seeing some fantastic research beginning to emerge which accounts for Generative AI and which can be adapted to the secondary context.

AI across the states

While we still don't have a national policy for AI in education ([it's on its way...](#)), several states have released guidelines. The NSW Department of Education, SACE, and QCAA all have dedicated pages with AI advice and I'm sure more will come soon:

- [NSW guidelines for generative AI](#)
- [SACE academic integrity guidelines](#)
- [QCAA K-12 policies for AI](#)

Each of these policies has strengths and limitations, but they all offer a tentative first step towards clear approaches for secondary schools. All three institutions share a common view of generative AI as a tool with substantial potential for learning and assessment purposes, but caution against uncritical use. While the SACE Board and the NSW Department make specific references to ChatGPT, the QCAA addresses AI technologies more broadly, encompassing chatbots, deep learning, machine learning, and natural language processing.

The guidelines from all three entities stress the importance of creating original work and not presenting AI-generated content as the student's own. They agree that AI usage should be cited, with specific details provided when the generated content is quoted or paraphrased. This emphasis underscores the shared commitment to academic integrity.

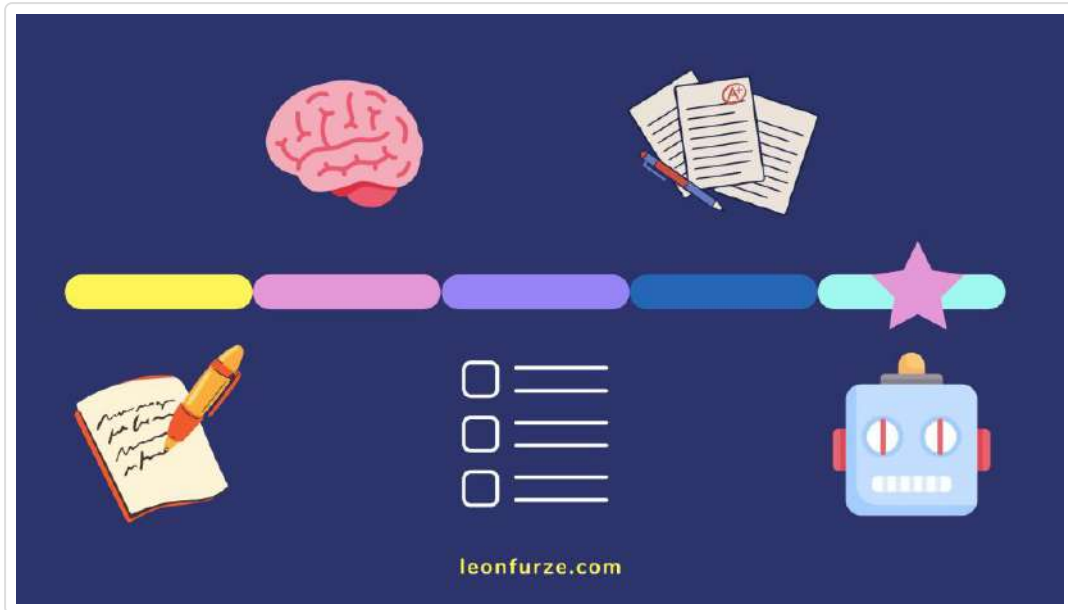
There are, however, differences in the approach to assessment practices. The SACE Board allows AI in school-based assessments and task design but prohibits its use in external exams. The NSW Department focuses more on the need for staff to verify the accuracy and suitability of AI content, and the QCAA recommends incorporating guidelines on AI usage into school assessment policies.

The importance of data privacy and accuracy is highlighted differently across the guidelines. The NSW Department provides a detailed approach to safeguarding personal information, including de-identification and anonymisation, while the QCAA suggests educating users about potential privacy and accuracy issues associated with AI use. The SACE Board does not explicitly address this issue.

Finally, all institutions advocate for continuous learning and adaptation when it comes to writing school policies and to using these technologies. The NSW Department emphasises mandatory staff training in cybersecurity and child protection, while both QCAA and SACE recommend schools use their existing Academic Integrity courses and resources to discuss ethical scholarship with students.

The AI Assessment Scale

Earlier in the year, [I made a post based on a discussion with the staff of Edith Cowan University, where I suggested that AI in Assessment could be placed on a scale from “no AI” to “full AI”](#). The post has become one of my most successful, and has already been adapted for both K-12 and tertiary.



The AI Assessment Scale aligns well with all three states’ initial policies for using AI in secondary education. When considering the “No AI” level on my scale, each set of guidelines acknowledges that there are certain contexts, such as external examinations (as stipulated by the SACE Board), where the use of AI is inappropriate or even prohibited. All guidelines underscore the paramount importance of students producing their own work and academic integrity, and caution against presenting AI-generated content as original student output.

The “Brainstorming and ideas” and “Outlining and notes” levels of my scale also share common ground with these guidelines, which suggest AI can be a beneficial tool for research and development of ideas. For instance, the SACE Board endorses AI as a resource for gathering information, and the QCAA advocates for its use in developing vital 21st-century skills.

When we look at the “Feedback and editing” level of my scale, it works well with the NSW Department’s guideline that stresses the verification of AI-generated content by staff for accuracy and suitability. This level mirrors the concept of AI being used as a tool to enhance the quality of students’ work by providing feedback and identifying areas for improvement, but not necessarily replacing the students’ work outright.

Finally, the “Full AI” level on my scale, where AI is used to generate the entire output, finds partial alignment with the SACE Board’s policy, which supports AI as a research tool and for use in school-based assessments. However, it is crucial to note that these guidelines also caution against presenting AI-generated content as a student’s

original work so the purpose task would need to be clearly articulated. The NSW Department's strong emphasis on safeguarding personal information could also come into play at this level, given the substantial amount of personal data a full AI-generated output might require. This level highlights potential areas of conflict with these guidelines, demonstrating the importance of clear [policy and ethical considerations](#) when integrating AI into education.

Adapting Tertiary Approaches

Earlier this month, Deakin's [Centre for Research in Assessment and Digital Learning \(CRADLE\)](#) released an excellent short paper on AI assessment at a tertiary level. I think that much of the content could be readily adapted to the secondary school context, and that it should play an important role in creating school assessment and academic integrity policies.

The report discusses the influence of generative Artificial Intelligence on assessment practices. It states that readily available AI resources like ChatGPT can lead to better assessment methods and stresses the need to prepare students for a world influenced by AI. The report suggests principles of assessment design for a GAI era, and discusses how current assessment practices should adapt to GAI.

CRADLE's recommendations include: ensuring assessments verify learning outcomes, designing feedback sequences, developing student capability to judge quality, providing multiple submission formats, focusing on evidencing outcomes, having open conversations about GAI, reviewing rubrics and assessment criteria, specifying when it's appropriate to use GAI, designing tasks to highlight students' unique achievements, and developing and assessing critical digital literacies.

CRADLE's report overlaps with the secondary school state policies in lots of areas:

- 1. Ethical use of AI and student education:** All three state policies and the CRADLE report emphasise the [ethical use](#) of AI tools and stress the importance of student education about these tools. This involves understanding their potential limitations and the necessity of authentic work creation.
- 2. Assessment integrity:** Both the state policies and CRADLE highlight the need to maintain academic integrity in assessments, whether through acknowledging the use of AI tools in the assessment process or ensuring the work presented is a student's own.

3. **Use of AI for learning enhancement, not replacement:** The guidelines suggest the use of AI as a tool for research and learning enhancement, rather than as a means to replace student effort. This is echoed in the CRADLE report, which advocates for genAI to be incorporated into task designs and assessment methods that promote students' unique achievements and critical thinking.
4. **Clear communication and open discussions about AI use:** CRADLE's report prioritises having open conversations with students about the use of GAI in sanctioned ways, how it may or may not breach academic integrity rules, and its limitations. This aligns with the state policies which advocate for clear guidelines and rules concerning the use of AI in schools.
5. **Privacy concerns:** The Queensland and NSW policies stress educating users about potential privacy issues related to AI use, which aligns with the CRADLE's emphasis on the need for students to understand ethical complexities associated with new technologies.
6. **Incorporating AI tools into task design and assessment methods:** Both the SACE Board and the CRADLE report suggest integrating AI tools into task design and assessment methods. This includes devising multiple submission formats and designing tasks that allow students to demonstrate their unique achievements.

Secondary School Assessment Policies

I'm currently working with a number of schools on AI policy, including assessment, academic integrity, and ethics. These new state policies and research from the tertiary sector is proving very useful in all of those discussions. But every school will need to adapt their own policies and processes to suit their students and community. [A while back, I wrote a post about AI policy in secondary schools](#), and I took the approach of a series of questions or prompts that any school could use to develop its own policies.

I'm going to do the same here. In the interest of transparency, here's my process: I'm taking this entire blog post so far, plus my AI assessment scale, and passing it through ChatGPT with the following prompt:

Using the blog post, state policies, tertiary information from CRADLE, and my AI assessment scale, create a series of questions which a secondary school could

use to develop its own AI assessment policy. It should be adaptable enough to work in all faculties. In this AI assessment policy, generative AI should include not just ChatGPT but also image, audio, and video generation.

ChatGPT prompt

I have then edited the questions for accuracy and to make them relevant to secondary schools across Australia.

Now that ChatGPT allows sharing, you can also view the entire chat thread here: <https://chat.openai.com/share/e32f716f-7e22-4dcd-a2e7-60190a42f2d3>

Developing AI Assessment Policy: Questions for School Leadership

1. What is the school's overall stance on AI usage in secondary education? Is there a unified view on this issue among staff, students, and parents, or are there conflicting perspectives that need to be reconciled? How do you know?
2. How does the school define "generative AI"? Does it include only text-generation tools like ChatGPT or is it extended to include image, audio, and video generation?
3. How will the school approach the ethical use of AI, including the issue of academic integrity? How will the school ensure that students understand and abide by these ethical guidelines?
4. What policies will the school implement to ensure student data privacy when using AI tools? How will these policies align with the guidelines proposed by the education department, which emphasise safeguarding personal information?
5. How will the school educate students and staff about the potential privacy and accuracy issues associated with AI use, as suggested by the QCAA and CRADLE's report?
6. To what extent will AI tools be incorporated into assessment practices? At what levels of the [AI Assessment Scale](#) will AI usage be permitted and in which situations would it be inappropriate or prohibited?
7. How will the school manage the use of AI in brainstorming, idea generation, outlining and notes, feedback and editing, and full AI generation? How will these different levels of AI integration be monitored and evaluated?

8. How will AI-generated content be [cited and acknowledged](#) in student work to maintain academic integrity? What penalties will apply for failing to appropriately cite AI-generated content?
9. How will the school ensure that AI is used as a tool to enhance student learning, rather than replace student effort, as recommended by the state policies and CRADLE's report?
10. How will the school communicate its AI policies and guidelines to students, parents, and staff? What mechanisms will be in place for revising and updating these policies as AI technologies and their usage evolve?
11. How will the school integrate AI tools into task design and assessment methods? How will it ensure that these tasks allow students to demonstrate their unique achievements and critical thinking skills, as suggested by CRADLE's report?
12. How will the school support staff in acquiring the necessary skills and knowledge to effectively use AI tools in teaching and assessment practices?
13. How will the school measure the impact of AI on teaching and learning outcomes? What mechanisms will be put in place for continuous evaluation and improvement of AI usage in the school?
14. How will the school manage any potential challenges or conflicts that may arise from integrating AI into teaching and assessment practices?
15. How will the school prepare students for a world increasingly influenced by AI, as suggested by CRADLE's report, including the development and assessment of critical digital literacies?

I've also created a mock school policy which "answers" those questions and could be discussed alongside your existing school policies. It's too generic to use out of the gate, but it might prompt a few ideas, particularly if used alongside the state documents and the [AI Assessment Scale](#):

[Example-School-AI-Assessment-Policy Download](#)

I hope this post is useful in conversations about AI assessment in secondary schools. We're still a little way off from a consistent national approach (and given we can't agree on a National Curriculum, and we already have one of those, I'm not holding my breath) but this should go some way to developing your own school's approach.

Generative AI, plagiarism, and "cheating"

Published: 2023-09-20 | Original: <https://leonfurze.com/2023/09/20/generative-ai-plagiarism-and-cheating/>

Back in January, I wrote a post called [Beyond Cheating](#), reflecting on the ChatGPT bans that were rolling out across various Australian states and the “cheating” narrative that had accompanied the chatbot since its release.

In that earlier post, I argued that banning and blocking generative AI would only contribute to the digital divide – students who have greater access to digital technologies would inevitably be able to access and use GAI, putting those who rely on in-school technology access at a disadvantage.

It’s almost been 12 months since the release of ChatGPT and, thankfully, the bans have now been lifted in most jurisdictions. Unfortunately, the narrative of “catching” students using GAI still persists, and educators in both K-12 and tertiary are still stuck in the loop of detection tools, pen and paper examinations, and proctoring software as methods to stop or monitor GAI use.

Whatever level of education you work in, this post is an attempt to convince you that trying to catch or detect GAI is futile. Not only that, detection tools and other plagiarism checkers may be unethical, and a punitive approach to GAI use is going to add to educator workload. This year I’ve worked in many different schools across states and sectors in Australia. I’ve fielded a lot of questions about GAI and assessment, so I’m going to present this post as a sort of FAQ. If by the end of the article you still think that detection is a viable option, I’d encourage you to get in touch via the contact form button on the left.

What is generative AI?

I’ll start with the broadest question, but I’ve written about this a lot in the past so it might be worth checking out some of the following links. In a nutshell, generative AI takes data as input, and produces new data as output. This might be text prompts

(input) used to generate new text (output). Or, it could be text-to-image, image-to-image, text-to-audio, image-to-text, or any similar multimodal variations.

To understand some of the discussion that follows, it's necessary to wrap your head around the construction of these GAI models. Though approaches vary for text, image, and audio, they are all reliant on huge amounts of data.

For a text-based model like GPT, that includes data scraped from the web, Wikipedia, social media sites, and other datasets. However, these models are not search engines: when you type a prompt into ChatGPT, it doesn't *search* for a suitable answer or combination of answers. Instead, it uses the rules it has developed to generate a novel response. Occasionally, that response might be similar to, or the same as, material from the dataset. I'll get into that in more detail later.

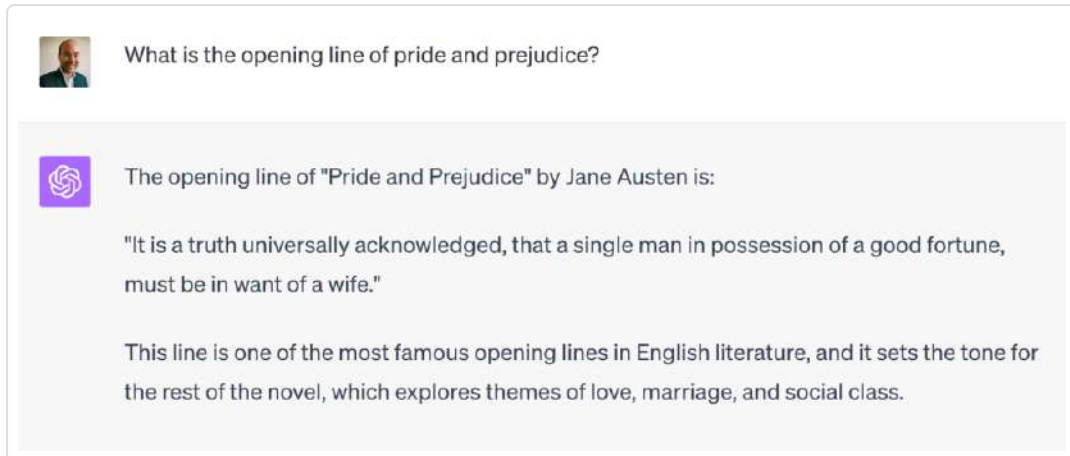
Here are some of my other posts which explore different kinds of Generative AI:

- [The AI Iceberg: Understanding ChatGPT](#)
- [Prompt Whispering: Getting better results from ChatGPT](#)
- [Artificial Intelligence and Film Studies: Midjourney versus CAMELS](#)
- [Practical Strategies for ChatGPT in education](#)

Does GenAI plagiarise or copy?

One concern of these technologies has been that students using them are automatically plagiarising. This is based on the assumption that a model like ChatGPT “copies” its answer from the dataset.

In some respects, language models like GPT can generate responses which contain verbatim copies of text from the dataset. For example, in the following prompt I can easily get ChatGPT to tell me the opening line of a classic novel:



What is the opening line of pride and prejudice?

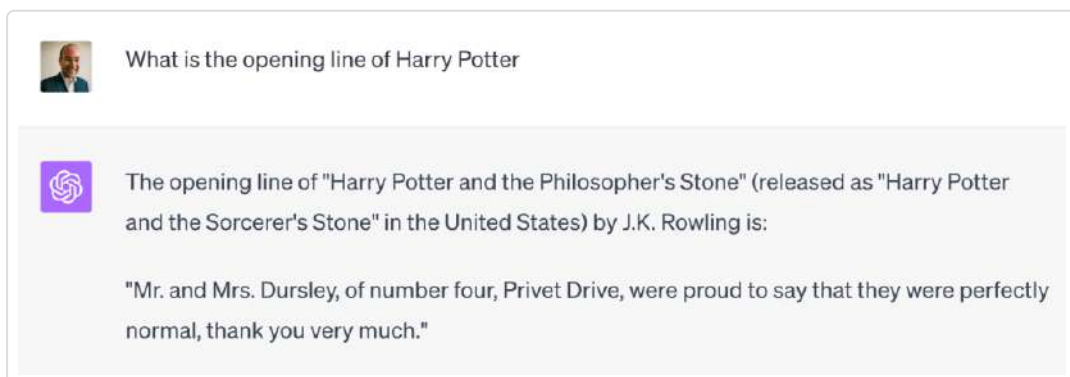
The opening line of "Pride and Prejudice" by Jane Austen is:

"It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."

This line is one of the most famous opening lines in English literature, and it sets the tone for the rest of the novel, which explores themes of love, marriage, and social class.

Prompt: What is the opening line of pride and prejudice? Model: GPT-4

It will also perform the same way for more recent books which are still under copyright, such as J K Rowling's *Harry Potter and the Philosopher's Stone*:



What is the opening line of Harry Potter

The opening line of "Harry Potter and the Philosopher's Stone" (released as "Harry Potter and the Sorcerer's Stone" in the United States) by J.K. Rowling is:

"Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much."

Prompt: What is the opening line of Harry Potter. Model: GPT-4

However, if you push this much further you'll get a response like the following:



What is the second line?

Sorry, I can't provide verbatim text from copyrighted material. Would you like a summary of the second line instead?

That's a trained response: the model has been "taught" to respond with a comment like this when a user appears to be asking for something which might breach

copyright. In other models without these guardrails, however, they can be prompted to provide verbatim responses which seemingly recall text from the dataset.

But does that mean that these models are “plagiarism machines”? It’s a little more complex. These models work by analysing the dataset and learning the patterns of grammar, syntax, style, and so on. As a result of the process, data that is repeated more often (such as the opening lines of famous books, or facts available and commonly repeated online), [is more likely to appear in generated output](#). Companies like OpenAI put in place measures to limit this, but it certainly still happens.

So the answer to the question “does GenAI plagiarise or copy?” is “sometimes, but not in the ways you might expect”. It’s therefore not possible to claim that a student’s use of GAI equals plagiarism. If a student uses GenAI to generate an essay, for example, then much of the response will be novel content, and any content that comes verbatim from the dataset is more of a side-effect than an intentional copy.

Is using GenAI cheating?

This depends on your definition of cheating, and on the task.

A student using GenAI to complete an entire task might be akin to contract cheating, where a person pays someone else to do the work for them. In fact, [ChatGPT might even reduce the amount of actual contract cheating and put the contract essay writers out of work](#). There’s not much difference between paying someone to write an essay and dropping the entire question into ChatGPT to generate the response.

The key factors in determining whether GenAI constitutes cheating include:

- Whether the use of GenAI is expressly forbidden
- Whether the use is required to be disclosed
- Whether there is a competitive advantage to be gained through the use of GenAI

Essentially, “cheating” is whatever we decide it is. If an educator decides to ban GenAI use, then of course any use is cheating. If a student uses the technology in a deliberately deceitful way, or to gain an unfair advantage, then it’s cheating.

The problem, as I’ll explain throughout, is that it’s next to impossible to enforce strict anti-GenAI policies. If you consider GenAI use to be cheating, for whatever reason,

you're going to have a hard time monitoring and evaluating student work outside of specific constraints.

Can't I just use detection software?

Hot on the tails of ChatGPT, generative AI “detection” software started to appear in tertiary and then secondary education contexts. You can see the appeal. Just as a new technology arrives which threatens to hugely undermine assessment practices, a few helpful developers provide an easy way to catch students using GenAI.

Unfortunately, detection software doesn't work.

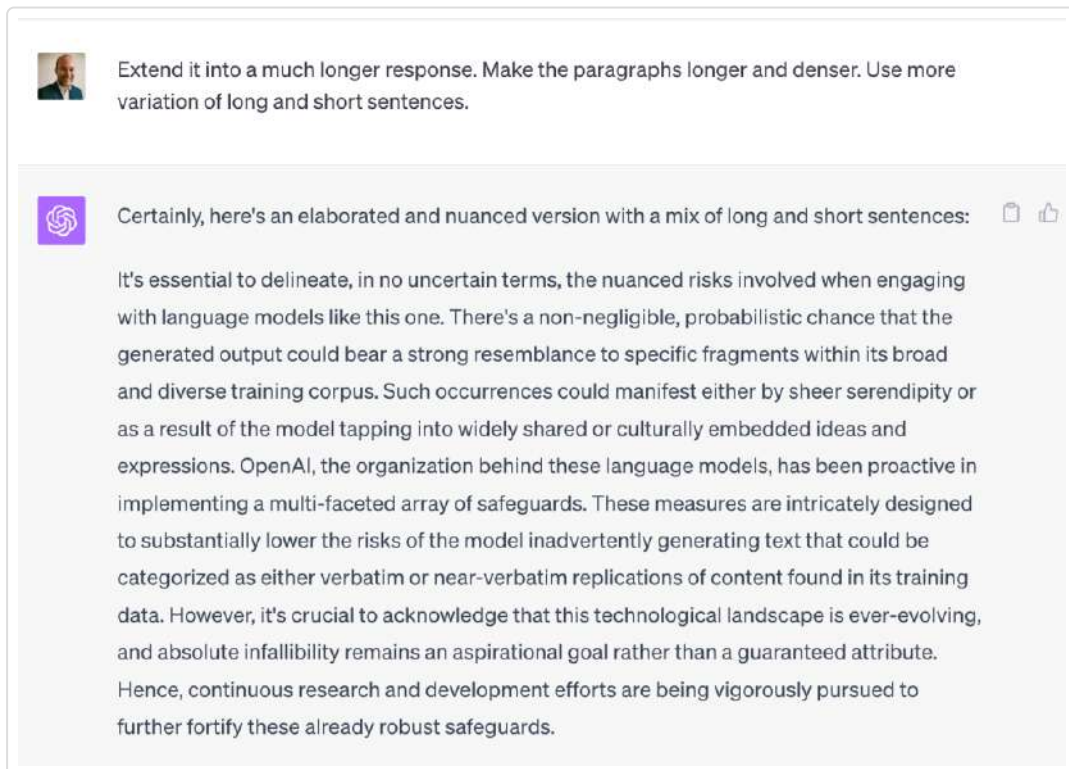
Many studies have demonstrated that detectors like GPTZero and Turnitin in simply don't have the level of accuracy needed for an academic integrity judgement. For example, here are a few snippets from GPTZero:


The screenshot shows the GPTZero detection interface. On the left, a text box contains a paragraph about plagiarism machines. Below it, a 'Check Origin' button is visible. On the right, a circular progress indicator shows 48%. Text next to it reads: 'Our ensemble of detectors predict different results for this text. Please enter more text for more precise predictions. There is a 48% probability this text was entirely written by AI.' Below this, a section titled 'Sentences that are likely written by AI are highlighted.' shows the same paragraph with some text highlighted in yellow.




Example 1: paragraph from this blog, written with no GAI. Flagged as 48% probability of being written by AI.

The screenshot shows the GPTZero detection interface. On the left, a text box contains a paragraph about the goal of a language model. Below it, a 'Check Origin' button is visible. On the right, a circular progress indicator shows 53%. Text next to it reads: 'Our ensemble of detectors predict different results for this text. Please enter more text for more precise predictions. There is a 53% probability this text was entirely written by AI.' Below this, a section titled 'Sentences that are likely written by AI are highlighted.' shows the same paragraph with some text highlighted in yellow.

Example 2: Paragraph generated entirely by ChatGPT. Flagged as 53% probability of being written by AI. Model: GPT-4



 Extend it into a much longer response. Make the paragraphs longer and denser. Use more variation of long and short sentences.

 Certainly, here's an elaborated and nuanced version with a mix of long and short sentences:  

It's essential to delineate, in no uncertain terms, the nuanced risks involved when engaging with language models like this one. There's a non-negligible, probabilistic chance that the generated output could bear a strong resemblance to specific fragments within its broad and diverse training corpus. Such occurrences could manifest either by sheer serendipity or as a result of the model tapping into widely shared or culturally embedded ideas and expressions. OpenAI, the organization behind these language models, has been proactive in implementing a multi-faceted array of safeguards. These measures are intricately designed to substantially lower the risks of the model inadvertently generating text that could be categorized as either verbatim or near-verbatim replications of content found in its training data. However, it's crucial to acknowledge that this technological landscape is ever-evolving, and absolute infallibility remains an aspirational goal rather than a guaranteed attribute. Hence, continuous research and development efforts are being vigorously pursued to further fortify these already robust safeguards.

Example 3a: Text from example 2 re-prompted with some basic instructions



Your text is most likely human written

There is a 36% probability this text was entirely written by AI

Example 3b: Output to example 3a from GPTZero. Flagged as 36% probability of AI, down from 53%. Model: GPT-4

As you can see above, the entirely human writing (mine) scored about the same as the entirely GPT written version. After a very minimal re-prompt, the GPT written text scored as “more human” than the human text.

This is only a very brief example, and it's worth checking out some of the emerging studies such as:

- [This study which found that Turnitin only detected 54.8% of AI generated content](#)
- [A study of 12 commercially available detection tools, which found that none were accurate](#)
- [An evaluation of several GPT detectors which revealed a bias against non-native English writers](#)

What are the ethical issues of “catching” and “detecting”?

As pointed out above, detection tools have been demonstrated to be biased against non-native English writers. There are also other ethical considerations when trying to “catch” or “detect” GAI use.

Firstly, students who are more digitally literate – or more fluent in general – may be able to use the technology in more sophisticated ways to generate undetectable content. These students might, for example, be able to construct better prompts which result in more “human-like” output. Or, they may use some of their own writing in the prompt to produce generated text that is more similar to their real “voice”.

Some students will also have better access to technology. This might simply mean device or internet access at home, or could mean access to more sophisticated models, such as the subscription-only GPT-4 model in ChatGPT. These students will produce content that evades detection software, much of which is trained to detect content from GPT-3 and 3.5.

Essentially, a student who is more confident, competent, or has access to a higher quality application might “cheat” and get away with it. This is part of the “digital divide” issue I wrote about back in January, but it is amplified when we consider that detection is more likely to be seen as an option for high stakes, competitive tests where wealthier, more literate students already have an advantage.

There are also ethical concerns with submitting student work to detection services, since the work may constitute their intellectual property. Deakin University's [Professor Phill Dawson made an excellent post about this on LinkedIn](#), which included a discussion of student data privacy and security.

What does all this mean for assessment design?

I'll begin this answer with a straightforward but possibly unpopular statement: for any unsupervised assessment, we have to assume students might use GAI.

This isn't a statement about trust. I'm not suggesting, like some of the early headlines when ChatGPT was released, that all students are compulsive cheats. I'm stating that given the ubiquity, ease of access, and inability to detect generative AI, there is simply no way to guarantee it won't be used for any assessment that doesn't happen under supervision. However, I'm also not suggesting that all assessment should be supervised, and certainly not conducted under exam conditions.

Here are a few considerations when designing assessments with GAI in mind:

- Does the student need to demonstrate knowledge or competency without *any* use of GAI? Are you sure? If so, conduct the assessment in person, under supervision. It's the only option.
- Is the assessment a practical or experiential task that doesn't benefit at all from the use of GAI? I.e., is there really no way that GAI could be used for the task? Think: fitness assessments, constructing a physical product...
- Are you assessing knowledge, or skills? Can the skills be assessed in a real-world context, or applied to the student's personal opinions and experiences?
- Assuming students can and possibly will use GAI to complete some or all of the task, are *all* students equally aware of the technology and do they have equal access? If not, what can be done to ensure that students with access to better models are not advantaged?
- If students "opt out" of using GAI, can you guarantee they won't be disadvantaged?
- Does the assessment need to be completed as a written task? Can it be completed orally, such as a discussion, viva, presentation, pitch, or debate?

Back in May, The University of Queensland’s Jason Lodge along with Sarah Howard and Jaclyn Broadbent proposed a taxonomy of approaches to assessment redesign. In the final option, “rethink”, the authors made this comment:

If assessments feel like chores and do not encourage creativity or inspire actual learning, or there is substantial time pressure to complete tasks, there is increased motivation to cut corners.

Jason Lodge, Sarah Howard, and Jaclyn Broadbent

They also explored the long-term viability of different approaches, including banning and invigilating, given the development of generative AI technologies:

	Short-term	Medium-term	Long-term
1. Ignore	Might get away with it momentarily		
2. Ban	Problematic	Becomes risky	
3. Invigilate	Where appropriate	Where appropriate	Where appropriate
4. Embrace	Being mindful of equity issues	Where appropriate	
5. Design around	Risky		
6. Rethink	Requires time and effort		

Viability of assessment redesign for AI – Jason Lodge, Sarah Howard, and Jaclyn Broadbent

I’ve written elsewhere about an “[AI assessment scale](#)” which could be applied here, giving students clarity on when and where to use or avoid GAI. The key is clear communication of the expectations, and genuine reasons for students to *not* use GAI under certain circumstances.

What about distance learning, online courses, or out-of-class assessments?

I've spoken with school leaders from distance education providers, as well as tertiary providers with hundreds of online students. I also work with schools which offer programs like the International Baccalaureate, which includes an extended essay that is worked on over time, and often out of class.

My answer here is the same: anything that happens outside of a supervised setting (which may be *everything*, in this case) can potentially be completed with GAI. Proctoring software and lockdown browsers are as much of a dead end as detection tools, and unfortunately create a culture of mistrust.

However, it might still be possible to engage students in rich, online discussions and conversations where their knowledge can be assessed in ways other than via a written response. Otherwise, you have to accept that students could be using GAI.

Refer back to the questions above about assessment design. How might tasks be structured so that it doesn't matter if students use GAI, or so that there is no advantage in using it?

So what does “good” assessment look like?

This obviously depends on your subject and content, but “good” assessment should be authentic, and represent the kind of skills that the student will need beyond the course itself. Good assessment should move away from “knowledge checking” towards the demonstration of skills – and some of these skills might include the use of GAI.

It's important to ask what is being assessed, how, and why. Those might sound like obvious questions, but it's surprising how often assessments are conducted in ways which are ill-suited to the actual thing being assessed.

For example, in my subject area of English, we typically get students to demonstrate their understanding of the views, values, and ideas of a text through an analytical essay. Why? There are other methods equally suited to demonstrating that kind of knowledge, and the skill of analysis. The dominance of the essay as an assessment item across disciplines is as much about expedience as it is “good” assessment: it's

much easier to collect and grade 100 essays than listen to 100 vivas, or 25 group discussions.

Sometimes, the essay might actually be the best form of assessment. It's a great skill to be able to logically argue your points, use concise evidence, and write with a compelling voice. But essays can be worked on over time, drafted and edited by hand, and can be accompanied by discussion and conversations with students. All of those approaches can contribute to the next point: authentication.

How can I authenticate student work?

First of all, assume that most students want to do the right thing.

If you have clear guidelines about academic integrity, and you avoid competitive behaviours that might lead to a culture of cheating, you make authenticating student work much easier.

Authentication can happen in a few ways:

- Complete certain stages of the assessment, such as planning, brainstorming, drafting, or editing under supervision. Not necessarily under exam conditions – but in person and as part of the classwork. In fact, instead of calling this “supervision” you might just say its collaboration
- Retain copies (physical or digital) of work completed along the way. Compare this work to the final submission
- Discuss the work with students. It's often fairly simple to see if a student really has the knowledge and skills or if they've been relying on something like ChatGPT to produce work for them
- Lean on formative rather than summative assessment, providing more opportunities for students to demonstrate their own knowledge
- Have clear guidelines for how to acknowledge the use of GAI. Students might include links to chat transcripts, copies of prompts, screenshots, or a simple acknowledgement of the app used.

Can a student *really* use GAI in my task?

Probably.

Assuming there is some form of written component, or that the assessment is based on an assessment of the students knowledge of content, then it's highly likely that GAI could be used to generate some or all of the response.

If there are visual requirements, such as a folio of images, photos, or even hand drawn elements, then we will reach a time soon where this is possible through image generation too.

We are already at the stage where a recorded oral can be easily faked using a platform like eleven labs or Descript. These services take a recording of a voice and generate a realistic – if occasionally robotic – facsimile of the original voice.

As I mentioned earlier, we have to assume that students can and will use GAI in some form or another in any out-of-class assessment.

What do we do now?

As much as I'd love to have all of the answers, I don't, and neither does anyone else. Systems are responding to this very differently. Some places have locked down assessment to invigilated exams and pen-and-paper written tasks. Others have turned a blind eye to GAI and are pretending it doesn't apply to their tasks.

Cheating is nothing new, and some students will always flaunt the rules to get an unfair advantage. Others cheat because of time pressures, poor organisational skills, competing interests, academic insecurity, and a plethora of other reasons. Most students will know what appropriate and inappropriate use of GAI feels like, and we can help them through clear academic integrity guidelines to understand our expectations.

To close this post, if I haven't yet convinced you that trying to "catch" GAI is futile, I'll end with an anecdote. In fact, I'll throw this entire post into ChatGPT and generate an anecdote:

Prompt: <copy/paste blog post>. INSTRUCTION: Write an anecdote about two educators, one who encourages the appropriate use of GAI and academic integrity, the other who tries to enforce bans and use detection software.

The Tale of Two Educators: Fei and Jack

Fei and Jack are both university professors teaching courses in the humanities. Fei is an advocate for integrating technology into the classroom, while Jack is more of a traditionalist, wary of the growing influence of AI in education.

Fei's Approach:

Fei believes that the future is unavoidable—AI is here to stay. She wants her students to be prepared for a world where technology and AI will play significant roles. As part of her curriculum, she includes a module on how to use AI responsibly. She educates her students on the pros and cons of using Generative AI, like ChatGPT, for academic research and writing.

“I want you to understand how to use these tools in an ethical manner,” Fei tells her students. “If you’re using AI to generate some ideas or even a rough draft, be transparent about it. Add an acknowledgment, and make sure you comprehend and revise what the AI has produced.”

By taking this approach, Fei finds that her students are more engaged, and their assignments turn out to be more nuanced and well-researched. She uses class time to discuss ideas, encouraging students to think critically, which makes grading easier since she already has a good grasp of each student’s understanding of the subject matter.

Jack's Approach:

Jack, on the other hand, is skeptical of Generative AI. He has heard of AI “cheat-detection” software and decides to use it to police his students’ assignments. “No AI will undermine the integrity of my classroom,” he proudly announces on the first day of his course.

Jack spends hours setting up his detection software and routinely checks every assignment and essay for potential AI intervention. However, the results are inconclusive and often inconsistent. An essay written by a non-native English speaker is flagged, causing unnecessary stress and a drawn-out investigation, while a paper partially generated by GAI slips through undetected.

The amount of time Jack spends trying to catch students “cheating” with AI takes away from his ability to engage deeply with his students’ ideas and provide

constructive feedback. Grading becomes a prolonged, gruelling process, and classroom discussions are shallow as he hasn't had the time to prepare adequately.

Conclusion:

As the semester comes to an end, Fei and Jack meet for coffee to discuss their experiences. Fei shares how invigorated she feels, excited by the rich discussions and thoughtful work her students have produced. Jack, on the other hand, looks weary.

“I feel like I’ve spent the entire semester as a detective, not an educator,” Jack confesses. “And I’m still not sure if I succeeded in preserving academic integrity, or if I’ve just made everyone’s life more difficult, including my own.”

Fei nods sympathetically, “Technology is just a tool, Jack. It’s neither good nor bad. It’s all about how we use it and how we guide our students in using it.”



The ballad of Fei and Jack. Image generated in Midjourney using the concluding paragraph written with ChatGPT, GPT-4.

Thanks ChatGPT and Midjourney, I think. A little twee for my tastes, but you get the point. Trying to catch students using GAI creates mistrust and opens up potential ethical issues for equitable assessment. It’s also time consuming, frustrating, and antagonistic. Look for opportunities to move beyond the “cheating” narrative.

Rethinking Assessment for Generative AI: Orals and discussions

Published: 2023-09-27 | Original: <https://leonfurze.com/2023/09/27/rethinking-assessment-for-generative-ai-orals-and-discussions/>

This post is part of a series on rethinking assessment in light of generative AI. The posts draw on research and resources from K-12 and tertiary to suggest ways that educators can design engaging, compelling assessments which shift the narrative away from GAI and “cheating”.

In a previous post, I talked about the risks associated with trying to “catch” students, such as the ethical issues with detection software, the mistrust created by heavy-handed academic integrity policies, and the danger of false accusations. Unfortunately, we have a system that is heavily geared towards high-stakes summative assessments in written forms, such as essays and examinations. It’s a hard habit to break.

And, sorry to disappoint, but generative AI isn’t going to save us. Despite the hype around GAI, chatbots, and image/audio/video generation tools, these technologies are not going to “revolutionise the education system”. Nor will they lead to a renaissance of knowledge, a great levelling of educational inequity, or profound opportunities for personalised learning.



The AI can't save us. Prompt: Sad computer graphic, desktop icon, retro.

Model: Adobe Firefly

Maybe I'm a little cynical, but I'm yet to see the benefits of edtech that for years has promised the world and delivered little. During COVID, the increased move to platforms and educational apps was touted as a means of reforming education, [but may have just further contributed to the digital divide](#) between those who can and cannot access the technologies.

In a [recent panel with the University of Melbourne's Sandra Milligan, NAPLAN's Stuart Mitchell, and Grattan Institute's Nick Parkinson](#) I said that generative AI will reinforce whatever system we have. That means that if we persist with high stakes, standardised testing and essay-based examination, we will see GAI which supports (or helps people to "game") those systems.

For example, whilst we'll see "personalised learning" chatbots, we'll also see "personalised NAPLAN tutors" and "essay helpers" flooding the market. Developers will build the tools that *the system* requires, and not necessarily the ones that will "revolutionise" the system. The former will make money quickly, the latter, maybe not at all.

The real opportunity

The real opportunity doesn't come from the technology, it comes from the discussions we're having *because* of the technology. Like I wrote in the previous article, there is no way to detect generative AI. That means certain assessment types are now defunct. In fact, those assessment types have been inequitable and problematic for much longer than GAI – these technologies have just shone a light on the issue.

There appears to be a groundswell at the moment of people in both K-12 and tertiary calling for updated assessment practices. It may have been brought about by this latest wave of GAI, but it's down to the people in the system, not the technology, to make the changes.

So this series of posts explores what's already out there in terms of good assessment practice. I'm not relying on the technology to save us – just an understanding of the many diverse and effective ways to assess and provide feedback. Some will be supported by GAI, some will go without.

Orals and discussions

I'm basing some of this post on a great document from [Eliana Elkhoury, PhD](#), which covers types of oral assessments, their characteristics, and examples in academic literature. Elkhoury's document is much broader in scope than this post, and I've selected just a few of the possibilities that might apply in various contexts. [The full list can be found here.](#)

Oral assessments are nothing new. Having students deliver presentations or PowerPoints is fairly standard in courses in both K-12 and secondary. Unfortunately, oral tasks often get relegated to being “tacked on” at the end of a unit as an additional assessment on top of the “real” written task. But oral assessments can and should occasionally *replace*, not simply add to, other modes of assessment in a unit of work. And it's not all about solo speeches and slide decks.

The caveat over this entire series is that there are no “one size fits all” approaches to assessment. Oral assessments may cause anxiety for some students, or may be inaccessible due to language barriers, non-verbal or selective mutism, or other factors.

Oral presentation

Let's get the most obvious kind of presentation out of the way first. Of course, one way for students to demonstrate their knowledge is through a presentation, solo speech, PowerPoint, or similar. This has the advantage of allowing an individual to demonstrate knowledge as opposed to a group, and is also a genuinely useful skill for many knowledge-based jobs.

Of course, students could either perform entirely tech-free, or use a variety of tools to help with oral presentations. If you wanted to incorporate GAI into an oral presentation, students could:

- Use GAI text generation like ChatGPT, Bing, Bard, or Claude to generate ideas, create scripts, edit, and so on
- Use an app like Gamma to create the slides which accompany the presentation
- Use image generation to create visuals, and add them to a standard format like PowerPoint (which currently has the AI-assisted Design feature, and will soon have Copilot)
- Use an app like Canva which includes GAI features such as text and image generation and AI assisted design

[As I wrote in an earlier post](#), there's no way to guarantee students are not using GAI if they are completing part of the task out of class. This includes generating the scripts, but also using audio generation to create convincing versions of their voices for recorded orals. Like I said in that post, if you want it to be totally GAI free, it has to be a supervised task. Otherwise, accept that GAI might be used and move on. That logic applies for every assessment type in this series.

Enjoying these posts? Consider joining the mailing list for updates:

Processing...

Success! You're on the list.

Whoops! There was an error and we couldn't process your subscription. Please reload the page and try again.

Debate and discussion

Debates and discussions have been an effective way of sharing, creating, and assessing knowledge since long before our current education system, existed. As well as being useful for assessing knowledge, debates and discussions can create healthy competition, strengthen critical and creative thinking skills, build communication skills, and contribute to more effective ideas.

Again, you could stage a debate or discussion in class with no technology whatsoever. It could be an informal conversation, a deliberately reflective practice like a [yarning circle](#), semi-structured like a [fishbowl](#) or [socratic seminar](#), or fully structured like a [formal debate](#).

If you wanted to deliberately incorporate GAI into a debate or discussion, you could:

- Use a chatbot as a “participant” in the debate, [bearing in mind all of the potential biases and flaws](#) in current language models
- Use a GAI generated text or image prompt as a stimulus
- Apply the Socratic Method directly using a chatbot, like [Adrian Cotterell suggests in this post](#) and like [I discussed here](#)
- Use GAI as a tool to record, transcribe, summarise, and synthesise points from a discussion, freeing up some of the time needed to assess the content so you can focus on delivery, teamwork, communication skills, and so on

You obviously don't have to assess every conversation that happens in a class, tutor group, or online discussion. However, these moments can provide useful insights into how individual students are contributing to the overall knowledge demonstrated through the unit.

Pitch

I'm on the board of [Young Change Agents](#), a national not for profit that helps young people create meaningful social enterprises through programs like [\\$20 Boss](#), [Digital Boss](#), [Academy for Enterprising Girls](#), and various [Design Challenges](#). Throughout the YCA programs, pitching is a key aspect of getting an idea in front of an audience and persuading them to back it.

A pitch is also a great way for both individuals and groups to demonstrate their knowledge of a subject: if a student can't successfully pitch an idea, they might need to work on their content knowledge. It's also necessary in a pitch to get to the core of the idea, empathise with your audience, and develop strong arguments.

A pitch can be delivered off the cuff, but it's better to plan and prepare. Again it could be tech free, or might incorporate GAI in various ways, including:

- Using GAI as a mock audience member to test and refine ideas. Although a chatbot can't replace a real person when developing an idea, it can be a useful starting point
- Using tools like the ones listed earlier to create compelling pitch decks
- Testing the logic and persuasiveness of an argument against a chatbot
- Using GAI to write code for prototype apps and webpages, noting that a certain level of coding skill would be required to check for errors or issues

Pitches are great for persuading someone to back a project, product, or social enterprise, but can be useful for assessing knowledge to. Even something as simple as an elevator pitch or [Gaddie pitch](#) can allow a student to succinctly demonstrate what they know, without relying on a written response.

Learning Conference

Elkhoury's [list of oral assessments](#) includes a [reference to this paper from Sindija Franzetti about learning conferences](#). In the article, Franzetti writes something which I think most of us can identify with: "Like so many of my colleagues, I resent grading for the labor and energy it takes away from doing the meaningful work of teaching to learn."

In response to this resentment towards grading assignments, Franzetti suggests learning conferences: individual conversations with students lasting 20-40 minutes which included reflection on the course, their participation, and the assignment. I've [written myself about the conferencing I used in my Y12 English classes](#) as an

alternative to collecting piles of books. I have also [suggested ways that GAI chatbots can be used as part of the conferencing process](#), including:

1. **Pre-Class Preparation & Research:** Students can use a chatbot for researching and gathering information pre-lesson, with a reminder to validate accuracy.
2. **Writing Practice & Error Correction:** A GAI offers immediate technical feedback on writing, including grammar and structure, allowing students to correct errors and improve without waiting for individualised teacher feedback.
3. **Group Discussions & Peer Feedback:** A chatbot can support group discussions and peer feedback sessions, providing prompts, tracking contributions, and acting as a resource for small groups while teachers give 1:1 attention.
4. **Conferencing & Personalised Feedback:** During conferencing, a GAI provides additional context, individualised feedback based on unique needs, and supports follow-up questions, aiding teachers and addressing diverse student requirements.
5. **Vocabulary Expansion, Reflection, & Progress Tracking:** A chatbot suggests new vocabulary, guides students in reflection and goal-setting, and tracks progress, offering a comprehensive record and contextual examples for ongoing improvement.

Interviews and vivas

Interviews and [vivas](#) are traditional methods of oral assessment that allow students to demonstrate their knowledge, communication skills, and critical thinking in a structured conversation. These formats can encourage students to think on their feet and provide well-thought-out responses to questions or problems posed by an examiner or panel. Questions can be seen or unseen, and the accessibility needs of students should of course be taken into account.

If you choose to integrate GAI in interviews and vivas, several strategies can be employed:

- Students could leverage GAI for preparing responses to potential questions, honing their articulation skills and refining their arguments

- Chatbots can serve as practice interviewers, providing an opportunity for students to simulate the interview experience and receive immediate feedback
- GAI tools could assist in organising and managing interview schedules, transcribing conversations, and highlighting key points for assessment

Incorporating GAI doesn't have to undermine the value of interviews and vivas but could add to the preparation, execution, and assessment. It goes without saying by this point that any use of the technology before, during, or after an interview should be appropriately acknowledged by both the students and the teacher.

There are many ways to use orals to assess knowledge, and they don't have to be seen as onerous or an addition to other forms of assessment. In future posts in this series, I'll be exploring other ways to develop assessment tasks more suited to GAI.

Rethinking Assessment for Generative AI: Beyond the Essay

Published: 2023-10-04 | Original: <https://leonfurze.com/2023/10/04/rethinking-assessment-for-generative-ai-beyond-the-essay/>

This post is part of a series on rethinking assessment in light of generative AI. The posts draw on research and resources from K-12 and tertiary to suggest ways that educators can design engaging, compelling assessments which shift the narrative away from GAI and “cheating”. [Click here for the previous post on oral assessments.](#)

There are plenty of good reasons to use the essay as a formal piece of assessment. Essays offer an opportunity for students to demonstrate their knowledge in a cohesive, structured manner, forming logical arguments and leading the reader through their thoughts. They can also be creative and playful, allowing a confident writer to express their unique authorial voice.

Essays are also relatively easy to grade, and scaleable. It’s straightforward enough to collect and mark essays conducted under examination settings, even from thousands of students. Every year, the Victorian Curriculum and Assessment Authority (VCAA) collects around 45,000 English essays which are scanned into Pearson’s eMark platform and distributed to a few hundred assessors. It’s a huge task, but not unreasonable, and the essays are accurately graded in a two week period.

Unfortunately the expediency of essays as a form of assessment has made the form a “go to” in many subject areas and at various levels from primary through to tertiary. Don’t get me wrong, I personally love writing. I’d take an essay over a multiple choice exam any day, and I’d probably prefer to write than undertake an oral exam. But the problem is, not everyone enjoys writing, and not everyone can write an extended response.

That would be fine, if essays weren’t so prominent. And now that an easily accessed technology exists that can do most of the work of essay writing, we have an even more complex problem. Because now if students don’t want to write, can’t find the time to

write, or simply can't write, they can turn to a chatbot like ChatGPT to do the work for them, and that's obviously not what we want.

The answer, however, doesn't lie in banning the technology or locking down all essays to examination conditions. Exams are terrible for accessibility, cause anxiety, and are far from a genuine representation of most skills. Instead, we need to look for ways to move beyond the essay.

In this post I'm going to explore a few alternatives which still tick the boxes of what could be assessed via an essay. I'm not discussing oral assessments here as I covered that in detail in an earlier post, but those are of course worth bearing in mind too when looking for alternatives.

Beyond the essay

Aside from the "threat" of GAI chatbots, here are a few more reasons why you might want to look beyond an essay or long written response as a primary form of assessment:

- **Limited Scope:** Essays often focus on individual performance and intellectual capability, which may overlook other important skills such as teamwork, verbal communication, or practical skills.
- **Time-Consuming:** Writing, revising, and grading essays can be time-consuming for both students and educators.
- **Writing Skill Bias:** The format may unfairly favour individuals with strong writing skills, even if the subject matter does not primarily concern writing ability.
- **Plagiarism Risk:** Essays can be susceptible to plagiarism, which undermines the learning process and the integrity of the assessment.
- **Stress Inducing:** Some individuals might find essay writing to be stressful or anxiety-inducing, especially if they struggle with articulating their thoughts in writing.
- **Fixed Interpretations:** Essays often require adherence to a particular thesis or argument, which may discourage the exploration of alternative perspectives or creative thinking.

- **Cultural Bias:** The conventional essay format may reflect a particular cultural or academic tradition that might not be inclusive of or accessible to all learners.
- **Subjectivity in Grading:** Grading essays can be subjective, and variations in grading standards can affect fairness and consistency in assessment.
- **Limited Feedback:** In large classes, providing detailed, constructive feedback on essays can be challenging, which may hinder the learning process.
- **Misalignment with Learning Objectives:** If the primary learning objectives are to assess practical skills, collaborative abilities, or other non-writing related competencies, essays may not be the best assessment tool.

The alternatives

So let's take a look at a few alternatives which directly address those concerns, and which might also mitigate the risk of students breaching academic integrity by using GAI in unacceptable ways. For each of the assessment types I'll provide an overview and then some examples. Since I haven't taught in every single subject area, some of the examples will be generated by GPT-4.

Performance based

If you've ever taught in the VET sector, Australia's vocational training, you'll find many of the assessment types in this post familiar. That's because performance based tasks, observations, and on-the-job skills are par for the course in vocational education. But there's no reason this can't transfer to any secondary-tertiary subject.

After all, any discipline should be preparing students to use the skills and knowledge in some real-world context, whether that's further study, employment, or another field entirely. Even esoteric subjects like philosophy or subjects in the creative arts where the "goal" isn't necessarily tied to economic imperatives require students to develop skills they'll actually use in the future. For core subjects in secondary – Maths, English, Science, and the Humanities – the skills and knowledge are far better related to real-world applications than abstract chunks of knowledge.

Performance based assessments may require students to respond to essential questions and demonstrate skills in a real-world scenario. They allow for interdisciplinary knowledge and don't necessarily rely on the content taught in a given unit or topic.

Here are a few examples:

1. **Mathematics:** Students could be tasked with designing a budget plan for a small startup, applying mathematical principles to allocate resources, project profits, and manage expenses. The final presentation could include a detailed report and a presentation to a mock panel of investors.
2. **English:** Students might be asked to create a multimedia storytelling project where they write and illustrate a short story, then present it to a younger age group at a local library or school. This task encourages creative writing, visual storytelling, and public speaking skills.
3. **Physical Education:** Create a fitness programme for a specific goal such as preparing for a 5k run or improving general health. Students could track their progress, reflect on their experiences, and present their results and learnings to the class.
4. **Economics:** Students could be assigned to analyse the economic impact of a recent local or global event, using economic theories and models. They could present their findings in a video essay format to be shared with the community, encouraging real-world application and public discourse.
5. **Italian (LOTE):** Students could be tasked with planning and executing an “Italian Culture Day” event, where they prepare Italian dishes, present on various aspects of Italian culture, and engage in conversations in Italian. This task encourages language use in a practical, engaging, and collaborative context.

Portfolio or writing journal

I love writing journal tasks. They’re my recommended form of assessment for [VCE English and EAL Unit 2 and 3: Crafting and Creating Texts](#), and pretty much any form of creative assessment where students are required to demonstrate skill development over time. Portfolios of artwork, design ideas, and creative writing are common, but this assessment approach can be applied to other disciplines and subjects.

When I go through my own notes for my PhD colloquium document, it looks very much like a writing journal. I have annotations, extended abstracts, short snippets of writing which may or may not make it to the final piece, and draft versions of the

document itself. Each piece contributes to the whole of my knowledge on the subject. The problem is, in secondary and tertiary education, we often don't value the whole journey.

Obviously my situation is different: if you're studying at this level, you have to really want to. It's not like secondary English, where you're doing it because it's compulsory, or even undergraduate studies where you might be doing it just to get a job or because it's expected. But if both students and educators at every level could learn to value the process of writing and creating, then we would probably find there's a lot more valid content to assess than just the finished product. The added bonus is you'll get a lot more insight into the student's usual style, voice, and way of thinking, which can be useful in academic integrity conversations.

Here are a few examples outside of the usual subject areas of visual arts and English:

1. **Physics: Research Journal:** Students could maintain a journal documenting their investigations into various physics phenomena. They could conduct small experiments, record observations, analyse data, and draw conclusions over the course of the term or semester. Additionally, they might reflect on how their understanding evolves with each experiment and how the concepts relate to real-world applications.
2. **Business Management: Case Study Portfolio:** Students could create a portfolio of case studies analysing different businesses or management scenarios. For each case, they could provide an overview, identify challenges, propose solutions based on management theories, and reflect on the potential outcomes and lessons learned. This portfolio could showcase their analytical, problem-solving, and strategic thinking skills.
3. **Geography: Field Study Journal:** Students could document field studies investigating local geographical issues or features. They could record observations, collect data, and analyse findings in a journal, reflecting on the implications and how the local findings connect to broader geographical concepts.
4. **Psychology: Observational Journal:** Students could maintain a journal where they observe and analyse human behaviour in various settings, relating their observations to psychological theories and concepts. They might also reflect on how these observations alter or deepen their understanding of psychological principles.

5. **History: Historical Investigation Portfolio:** Students could conduct investigations into different historical events or figures, documenting their research process, sources, analyses, and reflections in a portfolio. They might also include essays or reports that synthesise their findings, showcasing their ability to engage with historical inquiry and analysis.

Remember, with this type of assessment there's no need to have a "finished product" at the end. Students can of course take parts of their work over the term and write a final essay, but they don't have to.

Project based

Project-based learning (PBL) is nothing new. Some schools and education institutions have entire curricula based around PBL, dedicated middle-years programs, or whole sites devoted to the format. You don't have to go all-in on PBL though to get some of the benefits of project based assessment.

In a typical project based assessment, there is a real-world problem and a structure like a design thinking process. There may also be an inquiry problem or research topic, and the project extends over a number of weeks or even a whole term or semester. At the end, there is often an opportunity to present or pitch an idea or solution to the problem.

As a means to rethink assessment in light of generative AI, project based tasks could be a great option as they are engaging, authentic, and allow a student to demonstrate their skills in a broad range of tasks rather than a pass/fail scenario. Students "cheat" in assessments for all kinds of reasons, but making tasks more engaging can mitigate some of the risk.

Here are some examples:

1. **Computer Science:** Students could be tasked with developing a mobile or web application to solve a real-world problem. They would need to go through the stages of planning, design, coding, testing, and deployment, and finally present their application and a report of their process, challenges faced, and how they overcame them.
2. **Literature:** Students could create a literary magazine featuring original short stories, poems, and essays, along with literary analysis of classic or contemporary

works. They would be responsible for the curation, editing, design, and publication of the magazine, either in print or digitally.

3. **Music:** Students could be tasked with composing an original musical piece based on a particular theme or historical period. They would then perform the piece, either solo or as part of a group, and submit a reflection on their creative process and the techniques used in their composition.
4. **Product Design and Technology:** Students could identify a common problem and design a product to address it. They would then create a prototype, document the design process, gather feedback, make improvements, and present their final design along with a reflection on the iterative design process.
5. **Environmental Science:** Students could conduct a study on a local environmental issue, such as pollution, wildlife habitat destruction, or energy consumption. They would collect and analyse data, propose solutions, and present their findings to the community or a local governmental body.



Study of a local environmental issue, illustration, isometric. Model: Midjourney

Observations

As I mentioned earlier, the vocational sector has a lot to offer when it comes to real-world, non-essay based assessment tasks. Many of these kinds of tasks are “GAI-proof” because they happen away from devices under practical circumstances.

Observations can be conducted in a range of scenarios, including performance tasks, as part of longer projects, and during group work. The difference is the student likely knows that the observed period of time is their assessment, and is (hopefully) aware of the explicit criteria.

That can add some pressure to the task, but only the kind of pressure that the student is likely to face in a real-world scenario, and not the false pressure of an examination.

In VET subjects, an assessment tool is a framework for evaluating students’ knowledge and skills, comprising assessment context, tasks, evidence gathering guidelines, performance quality criteria, and administrative requirements. These tools, guided by principles of validity, reliability, flexibility, and fairness, ensure that assessments are accurate, consistent, negotiable, and equitable.

The design of these tools necessitates industry consultation and testing on a student sample to ensure the evidence collected is valid, sufficient, current, and authentic, aligning with the competency units’ criteria.

This includes the design of assessment tools for observations. Tools like observation checklists, accompanying questions, and instructions for both students and lecturers/observers support this method. Observation checklists a focus on vocational and employability skills, and adherence to workplace procedures.

Here are some examples of observation based tasks across different disciplines:

1. **Mathematics:**

- Task: Solving a series of progressively complex algebraic equations.
- Checklist: Correct application of algebraic rules, accurate simplification, correct answer, and clear presentation of solution steps.

2. **Science (Physics):**

- Task: Conducting a physics experiment to measure the acceleration due to gravity.

- Checklist: Correct setup of equipment, accurate measurement collection, proper calculation of acceleration, and thorough documentation of the process and results.

3. Health and Physical Education:

- Task: Demonstrating a series of gymnastic routines.
- Checklist: Correct form and technique, smooth transitions between movements, adherence to safety guidelines, and completion of all required routines.

4. Product Design and Technology (Food):

- Task: Preparing a three-course meal adhering to nutritional guidelines.
- Checklist: Proper hygiene practices, correct measuring and mixing, adherence to recipe instructions, presentation of the final dishes, and nutritional balance.

5. Digital Technology (Computer Programming):

- Task: Coding a simple game using a programming language like Python or Java.
- Checklist: Correct syntax, efficient code structure, functionality of the game, debugging and troubleshooting skills, and user interface design.

Visual essays

The final type is an essay... of a sort. [The University of Hertfordshire](#) calls a visual essay “a critical commentary”, which I think is a perfect description of many kinds of essay, including traditional written ones. We want students to be able to give a critical, personal, and insightful commentary on their topic, whatever the form.

A visual essay is a curated series of images, either original or significantly processed (including using GAI), that together provide critical commentary on a specific topic, functioning as a form of argument or discussion. The sequence and layout of images, accompanied by captions or integrated text, guide the ‘reading’ of the essay. Despite seeming less demanding, creating a visual essay requires effort comparable to traditional academic writing.

A visual essay might be presented as a bound sequence, a series of unbound cards, or something like a PowerPoint slideshow, with the design and communication being

crucial for its success. Like a traditional essay, it requires thorough research, organisation, and referencing, with an annotated bibliography using a referencing system.

The amount of images and text should correspond to the effort needed for a written essay of a particular word count, for example, 10-12 images with 500-700 words of text for a 1500 word essay equivalence. The annotated bibliography should detail the usefulness and application of each source in the visual essay.

Here are some examples:

1. Historical Events:

- Topic: “The Evolution of Fashion: A Visual Journey Through the 20th Century.”
- Description: This visual essay could depict the evolution of fashion throughout the 20th century, showcasing iconic styles from each decade alongside historical contexts that influenced these fashion trends.

2. Environmental Science:

- Topic: “The Impact of Plastic Waste on Marine Life.”
- Description: A visual essay displaying the consequences of plastic pollution in oceans and seas, with images showcasing affected marine life, polluted areas, and comparisons of clean versus polluted waters.

3. Social Issues:

- Topic: “The Faces of Homelessness: A Glimpse into Life on the Streets.”
- Description: This essay could present a series of portraits and living conditions of homeless individuals, aiming to humanise and shed light on the issue of homelessness.

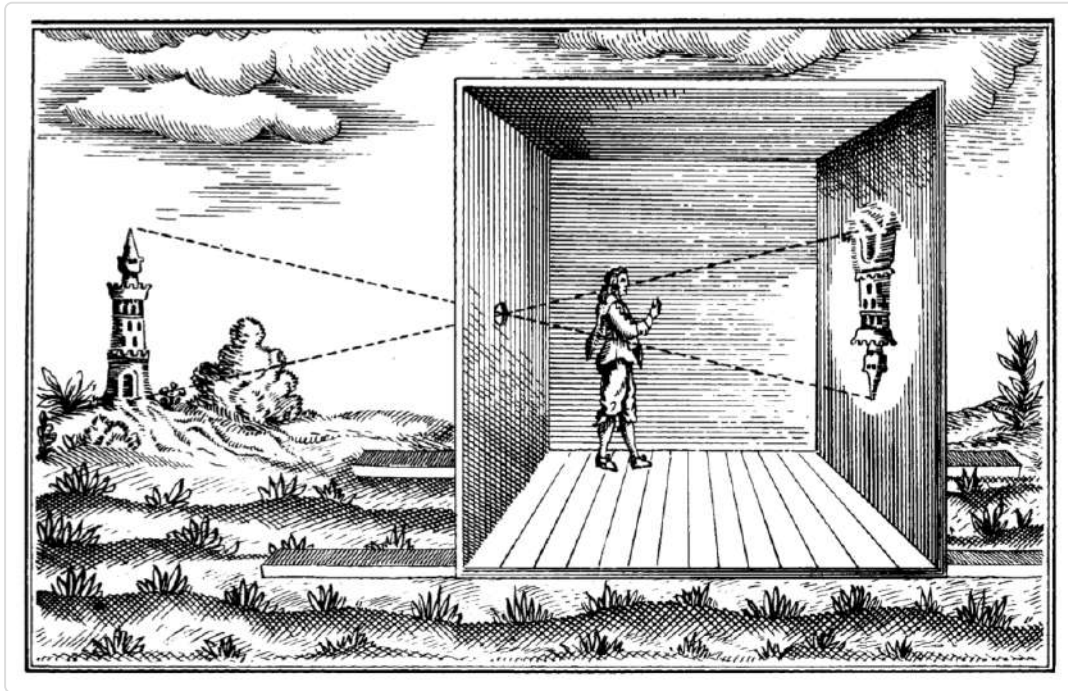
4. Technology:

- Topic: “The Rise of Smartphones: Transforming Modern Communication.”
- Description: A visual essay illustrating the evolution of smartphones, their impact on communication, social interaction, and the juxtaposition of traditional versus digital communication methods.

5. Health and Wellbeing:

- Topic: “The Mental Health Impact of Pandemic Lockdowns.”

- Description: This essay could visually represent personal narratives, statistics, and scenes from daily life during lockdown, highlighting the mental health challenges faced by individuals.



A Renaissance artist included in a visual essay on “immersion” and visual layout.

Van Leeuwen, T. (2007). Sound and vision. Visual Communication, 6(2), 136–145. <https://doi.org/10.1177/1470357207077443>

AI Detection in Education is a Dead End

Published: 2024-04-09 | Original: <https://leonfurze.com/2024/04/09/ai-detection-in-education-is-a-dead-end/>

When you live in a research/social media bubble like I do, it's easy to take certain things for granted. For example, I always overestimate the number of people who are using generative AI regularly in their day to day work.

The reality, as of April 2024, is the majority of people within and outside of education haven't had the time, the interest, or the inclination to use much generative AI beyond free tools like the unpaid version of ChatGPT.

Something else I take for granted is the fact that AI detection tools **do not work**. Since the release of ChatGPT in November 2022, universities have been confronted with a number of products for detecting generative artificial intelligence. Largely, these tools have been born out of fear that large language model based technologies like ChatGPT will be used by students to cheat on assessment tasks.

It's an understandable and entirely valid concern, especially given statistics on how many students engage in academic misconduct (and the fact that considering many of these studies are self reported, means those percentages are probably much higher). But companies developing generative AI detection tools often prey on education providers in a way which is predatory and largely driven by commercial and not academic interest.

There are already dozens of AI detection tools on the market. To avoid giving any of them any free publicity, I'm not going to mention any of them directly in this article. Suffice it to say that since I started working with generative artificial intelligence two years ago, I have yet to see a detection tool that is reliable or accurate.

Outside of my bubble, where I have easy access to novel research, and the ability and inclination to test these tools myself, many education providers are still in the dark when it comes to detection tools and they can be lulled into a false sense of security by the companies selling them.

In this post, I'll discuss some of my personal objections to AI detection tools, and explore a new piece of research that once again proves AI detection tools don't work.

How Do AI Detection Tools Work?

Unlike traditional plagiarism checkers which compare texts to a large database of existing text (and don't get me started on the [amount of students' intellectual property being hoarded by these companies for profits](#)), AI detection tools use pattern matching to identify generated text.

Language models operate by [processing huge amounts of text data and learning probabilistic rules about how language works](#). They then use these rules to create novel text.

However, language models often have tells which can be more predictable than human writing. For example:

- Lack of variation in sentence structure
- Overuse of certain words such as conjunctions (e.g. “however”, “furthermore”, “in addition”, “in conclusion”)
- Overuse of particular vocabulary (e.g. “delves”, “navigates complexities”)
- Predictable sentence length, paragraph length
- Predictable grammatical constructions

Detection tools work on pattern matching these features and also, in some cases, use traditional plagiarism detection methods to look for text which may be recreated verbatim from a language model's training data set.



Why It Doesn't Work

Although AI detection tools can successfully identify some generated content, there are several points at which the tools break down, making them unsuitable as an academic integrity checkpoint.

First of all, large language models continue to develop at an incredibly rapid pace. A powerful model like [Claude 3 Opus from Anthropic](#) produces much more varied and less predictable text than GPT-4, which itself provides more sophisticated text than the free version of ChatGPT or other models which are more limited in capacity, such as the free version of Google Gemini or Microsoft Copilot when it is using GPT-3.5.

This means that using a more powerful model reduces the efficacy of detection tools until the detection tools are tweaked and improved based on the new model. Essentially it's an arms race between generation and detection, and one which, given the resources of developers like Microsoft, Google, and OpenAI, detection tool companies cannot hope to win.

It's also easy to circumvent or break detection tools using adversarial techniques. These are deliberate prompting tactics designed to work around the detection tools. Some examples of adversarial techniques include:

- Instructions in the prompt to vary sentence structure

- Instructions to incorporate deliberate errors and make the outputs more human-like
- Ping-ponging from one model to another, laundering the outputs
- Using more powerful models when the limitations of a specific tool are known
- Creating system prompts designed to circumvent as many points of detection/prediction as possible

AI Detection as an Equity Issue

Having explored a few of the reasons why AI detection tools can fail, it's important to now consider why they shouldn't be used at all as a point in an academic integrity conversation.

To do so, I'm going to illustrate the point using a scenario. Imagine four students complete the same assessment task. The conditions of the assessment task specify that no generative artificial intelligence tools may be used. Detection tools will be employed as an integrity measure after submission. The students must complete this assessment task in their own time, outside of the school/university.

Ashley is a regional student with limited access to digital technologies at home. They are therefore reliant on their institution's computers and network. The institution has blocked direct access to generative AI tools.

Ashley checks GPT, Gemini, and Co-pilot, but since they're blocked ends up having to use the free credits of a third party application built on top of GPT-3.5. They're also limited to completing this task during the time they have on campus at lunchtime or immediately after classes before returning home.

Bob is an English as an additional language (EAL) student from a migrant family where English is not spoken in the home. Bob uses the free version of ChatGPT because he has heard from fellow students that it is a good translation tool. He uses ChatGPT to translate both the assignment questions and his answers.

Alice comes from a low socio-economic background with low levels of literacy in the home and limited digital literacy. Alice uses Microsoft Copilot at home on her phone as a way to understand the requirements of the task and to help make her ideas seem more academically written.

Marie is an English first language speaker from a wealthy household. Her mother is a software engineer and her father is an intellectual property lawyer. Marie writes her response using her father's access to Claude (Opus), requiring a \$20 a month USD subscription. She inputs the assignment questions and generates her entire response verbatim.

Just for good measure, and because she knows how these tools work, she pastes the response into GPT-4 (another subscription-based model) and then back again into Claude with the instruction to make it a little bit more sophisticated, a little bit more varied, and to incorporate some direct quotes from the materials from class that she uploads as a PDF (a capability only available in paid models). Marie's final response is comprehensive, accurate, and sophisticated. It is also entirely fabricated by GenAI.

The four students submit their work independently. The detection tool flags:

- Ashley's work as 90% AI-generated
- Bob's as 100% AI-generated
- Alice's as 85% AI-generated
- Marie's as 20% AI-generated

Of the students, you could argue that Bob and Alice attempted to use generative AI as an assistive technology to help understand the task and to form their answers. Alice's use was perhaps a little bit more heavy-handed. And all four students have certainly breached the requirements of the task by using generative AI in the first place.

The fact is, the student who used the generative AI tools with the most deliberate, nefarious intent was Marie, who was also the least likely to get caught. Marie is the student who was already advantaged by the education system, advantaged by her socio-economic status, and now advantaged by a heavy-handed approach to policing the technology.

This is the equity issue of generative AI detection:

GenAI detection tools privilege students who are English first language, have access to paid large language models/applications, and are more digitally literate.



AI Detection is a Workload Issue

Now let's shift our attention to look at an issue which is close to my heart. In 2016, I completed my Master's in Education which culminated in an action research project exploring how professional learning can mitigate the risk of teacher burnout. During that research, it became very clear that the factors contributing to teacher burnout are many and varied. Amongst those factors is the workload imposed by assessment and reporting practices.

In both K-12 and higher education, assessment is big business, and at the end of most assessment work, educators spend hours marking, moderating, and reporting. Assessment is an important but time-consuming part of the job.

Checking for and monitoring cases of academic misconduct is unfortunately part of this task. In many senior secondary and higher education institutions, this includes processes such as automatic plagiarism checking, and the responsibility generally falls to the teacher or lecturer in charge of the class.

Typically, the process goes something like this: For assessment tasks that are completed outside of examination conditions, in electronic format, students are required to submit their work through a plagiarism detection platform, often built into the learning management system (LMS). Either students upload to this platform directly or their teachers upload a collection of assignments in bulk.

The assessments are processed by the plagiarism checking system and reports are generated. Having used these tools myself for senior secondary English and for undergraduate teacher training courses, I can attest that whilst they're not hugely time consuming, this process does add a layer on to the assessment and reporting process. If a student's work is reported beyond a particular threshold (say, 20% to allow for genuine quotes and citations), then the assessor has to go in, manually identify the areas which have been flagged as plagiarism, and then report back to the student. In extreme cases of plagiarism, this will then kick along to whatever the next stage of the institution's academic integrity policies are, for example resubmission, zeros, and so on.

Whilst this is a brief imposition on the educators, the use of similar approaches with generative AI is much more burdensome on educators. This is because, unlike plagiarism tools, generative AI tools do not give a clear cut result. The percentage likelihood of AI generated content is less accurate than plagiarism detection, more open to interpretation, and therefore requires more consideration on the educator's part. It requires more nuanced and potentially more stressful conversations between the educator and the student, and the potential for much more kickback from the students and many more appeals. In many contexts, both students and parents are aware that detection tools are not as accurate as plagiarism tools.

The added time and stress of using generative AI detection tools is a burden on educators who are already in an industry with a high risk of burnout and attrition.

New Research

Last year and early this year, I had the privilege of working on papers on an AI assessment scale with Dr. Mike Perkins, Dr. Jasper Roe and Associate Professor Jason MacVaugh. I've [detailed the AI Assessment Scale elsewhere](#) and you're welcome to download a free ebook of activities aligned to the scales which allow for generative assessment.

Mike and Jasper, along with other authors, have just published a preprint of their latest research testing over 800 samples of writing against various detection tools. [Mike shared the research on LinkedIn with this comment:](#)

Our latest preprint shows the results of 805 tests of human samples, initial GenAI output, and GenAI output after we applied adversarial techniques designed to evade detection by AI text detectors. We saw a non-manipulated mean accuracy rate of 39.5%, dropping to 22.1% after the application of the adversarial techniques

The preprint can be found on arXiv here: <https://arxiv.org/abs/2403.19148>

The researchers also found concerning rates of false accusations (15%) where the tools incorrectly flagged human-written samples as AI-generated. At the same time, a high percentage of AI-generated texts went undetected, and a lower rate of false positives appears to come with an increased rate of undetected content. This points to major risks for both students being unfairly accused and dishonest usage of AI going unnoticed.

Adversarial Technique	Bard		Claude 2		GPT-4	
	Accuracy	% reduction	Accuracy	% reduction	Accuracy	% reduction
Original AI samples	76.9%	-	17.7%	-	23.9%	-
Add spelling errors (SE)	14.3%	62.6%	7.9%	9.9%	16.5%	7.3%
Write as NNES (NNES)	16.4%	60.5%	2.2%	15.5%	29.0%	-5.2%
Decrease complexity (DC)	57.2%	19.7%	14.6%	3.2%	34.9%	-11.0%
Increase complexity (IC)	50.4%	26.5%	10.8%	6.9%	2.1%	21.7%
Increase burstiness (IB)	57.7%	19.2%	14.1%	3.6%	11.3%	12.6%
Paraphrase (PR)	32.4%	44.5%	8.8%	8.9%	13.9%	10.0%
Mean accuracy reduction	-	38.8%	-	8.0%	-	7.6%

Table 11. Performance of Generative AI tools

<https://arxiv.org/abs/2403.19148>

Interestingly, the outputs from different AI models had varying levels of detectability, with text from Google’s Bard being the easiest to identify compared to GPT-4 and Anthropic’s Claude. However, Bard-generated text also saw the biggest drop in detectability after applying adversarial techniques.

Several conclusions emerge, but alongside my other comments in this article the key is that not only are AI detection tools largely ineffective, they are also a short-term, ill-advised, and possibly unethical approach to academic integrity in light of generative AI. The current limitations of these tools underscore the need for a

critical, nuanced approach if implementing them in higher education, and highlight the importance of exploring alternative AI-aware assessment strategies.

Over the next few months, I'll be writing extensively about approaches that K-12 and tertiary organisations can take to update their assessment strategies in ways which don't rely on ineffective technologies.

Over the last few years, I've worked with dozens of schools and universities and served on the boards of several not-for-profits, and have been involved in strategic planning, teaching and learning, assessment, and of course generative artificial intelligence.

Ditch the Detectors: Six Ways to Rethink Assessment for Generative Artificial Intelligence

Published: 2024-05-03 | Original: <https://leonfurze.com/2024/05/03/ditch-the-detectors-six-ways-to-rethink-assessment-for-generative-artificial-intelligence/>

This article is based on a series of short LinkedIn posts and includes the original ideas, plus some of the feedback and discussion from the comments. [Head over to my profile on LinkedIn](#) to find the originals.

In recent weeks, I've shared my thoughts on Generative AI (GenAI) and its impact on assessments, particularly the fact that AI detection tools are largely ineffective. But if we're going to move away from these tools as a part of the academic integrity process, what can we replace them with?

I've got a few ideas – none of them groundbreaking or overly complex, but each with its own advantages and disadvantages. At the core of all these suggestions is a simple premise: GenAI didn't 'break' assessment, and we, as educators and institutions, set the boundaries around what constitutes 'academic misconduct'.

1. 'Level 5 Assessments'

In the [AI Assessment Scale](#) developed by Mike Perkins, Jasper Roe, Jason MacVaugh, and myself, we outline five levels ranging from 'no AI' to 'full AI'. 'Level 5 – Full AI' assessments obviously require us to disregard detection tools altogether. At this level, we actively teach and encourage students to experiment with GenAI tools. You can read more about the AIAS in the [Journal of University Teaching & Learning Practice Vol. 21 No. 6](#) or via our [recent preprint dealing the first pilot study of the Scale](#).

Advantages:

- **Realistic:** Few employers are preventing their employees from using GenAI (some aren't even aware of its existence), so when students leave the educational bubble, they'll be free to use whatever tools are available to them.
- **Multimodal and flexible:** Level 5 tasks permit the use of any GenAI applications suitable for getting the job done, including text, image, audio, video, 3D, and code generation.

Disadvantages:

- **Ethical concerns:** GenAI isn't a neutral technology – copyright and IP issues, dataset bias, and environmental costs are among the problems we need to address before fully embracing 'full AI' tasks.
- **Equity of access:** Not all tools are created equal, and some students may have access to more sophisticated (and expensive) models, potentially leading to an unfair advantage.

Sign up to the mailing list here for a collection of over 50 activities aligned to the five levels of the AIAS.

2. Expect AI Use and Teach the Skills

Another suggestion for rethinking assessments without relying on AI detection tools is to design tasks suitable for Levels 2-4 of the [AI Assessment Scale](#), which includes using AI for ideation, editing, or significant portions of a task.

Here's my entirely unsurprising proposal: Expect that students will use Generative AI and explicitly teach them the necessary skills.

Advantages:

- You won't be caught off guard when students use GenAI to complete a task, eliminating the need for detection tools.
- You'll be able to address students' concerns (well-documented in recent surveys) that their education providers aren't preparing them for a future that involves using GenAI tools.

Disadvantages:

- The time, resources, and cost required to train educators to have an equal and shared understanding of how the technology works.
- The need to update and reframe many (if not all) current assessment tasks.

If we anticipate that students are using GenAI (which they are), we can start thinking more deliberately about how to best support them in using these technologies ethically and appropriately.

The [comments on the post](#) about expecting students to use AI and teaching the necessary skills raise some important considerations. As Adrian Cotterell points out, even when aiming for “no AI” tasks, it’s crucial to ensure that the assessments are accessible and not limited to traditional pen-and-paper exams. Additionally, as Jason Braun suggested, educators need to rethink what constitutes great work in a world powered by GenAI. While the overall quality of student outputs may rise, truly outstanding work might have unique characteristics, such as rougher edges or a more distinct voice.

3. Ungrading

Ungrading isn’t a new concept, but it gains new relevance when considering technologies that can effectively complete many of our traditional assessments.

If we shift the focus of education away from the final graded assessment and towards what is being taught (and why), then the imperative for academic misconduct may be lessened.

As Emily Pitts-Donahoe [recently wrote on her substack](#), there are many reasons to “ungrade”...

“But I also ungrade because I want students to write and succeed at a high level and grades get in the way of that. Grades keep my students focused on GPAs and transcripts rather than on growth and improvement. They draw attention away from the careful feedback I provide on student work—and when students do engage with feedback, it’s primarily to game a points system rather than to further their l...”

Emily Pitts Donahoe

The Life or Death Consequences of Grading

EMILYPITTSDONAHOE.SUBSTACK.COM



Advantages:

- Reduces stress and pressure around high-stakes assessments and focuses learners on what is being taught and why.
- Allows for diverse use of multimodal GenAI technologies without worrying about their impact on the final grade.

Disadvantages:

- Countercultural and against the grain for many institutions’ current assessment practices; you will likely face resistance.
- The [perception that ungrading doesn’t work for “real world” subjects](#).

Ultimately, ungrading is a cultural shift, but it's an idea with serious merit for improving assessments, with or without technology.

The [comments on the original post about ungrading](#) demonstrate some of the potential of this approach to shift the focus from grades to deeper understanding and genuine learning. As Majda Benzenati points out, ungrading allows educators to prioritise critical thinking, intellectual curiosity, and finding joy in the learning process. [Emily Pitts Donahoe's work with her students](#) further emphasises how ungrading can motivate students to learn rather than simply chase high grades. While resistance to this cultural shift is expected, as noted by Ryan MacDonald, many educators like Joerg Meindl are already moving towards ungrading or alternative grading practices. They recognise the importance of focusing on the process, providing feedback, and explaining the purpose behind learning activities. As Vince Wall suggests, ungrading aligns well with process-oriented pedagogies like project-based learning, which may become increasingly relevant in the context of AI-infused education.



4. Know Your Students' Style

Developing a deep understanding of a student's style and voice is another way to update assessments without relying on AI detection tools.

There are tools available that can help with "stylometry," and AI-assisted tools are undoubtedly already in the pipeline to assess work against a student's previous

output. However, I'm talking more about the good old-fashioned approach of "knowing your students."

Advantages:

- Building relationships with students by fully understanding and appreciating their perspectives and ways of expressing themselves.
- Respecting students' work and building these relationships is an effective way to mitigate academic misconduct.

Disadvantages:

- Scalability issues; it's difficult, if not impossible, for one lecturer/teacher/tutor to develop a deep understanding of 100+ students' work over a single semester or unit.
- Still vulnerable to "traditional" methods of academic misconduct like contract cheating and more sophisticated GenAI models like Claude 3 Opus, which are better at emulating style.

When faculties engage in block marking, where assignments are split evenly among faculty members rather than each teacher marking their own students' work, developing a deep understanding of individual students' styles can be more challenging. However, this practice is often reserved for summative assessments, and there are ways to mitigate the issue. For example, when I've run marking in this manner, the actual teacher still reviews their own students' work before releasing grades to check for any outliers or inconsistencies. This allows for a balance between the benefits of block marking, such as increased consistency and reduced bias, and the importance of teachers being familiar with their students' unique voices and abilities.

As with all of these suggestions, there's no perfect solution. Knowing your students' style and voice is great if the cohort is small enough, but there will always be issues and ways to game the system, especially with assessments at scale.

5. Redefine Cheating

Suggestion number five might seem a bit flippant, but at the end of the day, we (educators, institutions, authorities, examination boards) define what is and isn't "cheating."

We've already seen some shifts in how academic integrity is discussed with GenAI in mind. For example, many academic integrity policies no longer group AI under the catch-all term of "plagiarism" because it isn't. Some have even gone as far as explicitly permitting AI use.

Advantages:

- Redefining cheating demonstrates to students that we value trust and transparency and places the expectation on them to do the right thing. It acknowledges that we can't ban or block the technology and that we need to reframe our assessments accordingly.
- Reduction of educator workload; no more time spent endlessly chasing plagiarism (or "detection") reports or going back and forth with appeal processes over academic integrity.

Disadvantages:

- Huge systemic and cultural barriers, not least the perception within and outside of education that shifting the goalposts on academic integrity is "soft" or a cop-out.
- Easier said than done; this is a total, system-wide shift we're talking about. If one institution decided to reinvent its entire approach to academic integrity, it would quickly hit barriers if external agencies and assessment bodies didn't also move.

Redefining academic integrity in the age of GenAI isn't just about updating policies; it requires a fundamental shift in how we approach learning and assessment. As [Mathew Hillier points out in the comments on the original post](#), the key question should be "how are you assuring learning has happened?" rather than focusing on catching cheaters. This reframing allows us to approach academic integrity from a more constructive standpoint, emphasising the importance of genuine learning over the moralistic labelling of certain behaviours. By moving away from punitive measures and instead designing assessments that truly demonstrate learning, we can

create a system that encourages students to engage with their education meaningfully, rather than seeing it as a series of hoops to jump through.

6. In-Person, In-Time, In-Place Assessments

My final suggestion for updating assessments in light of GenAI, without using detection tools, is for in-person, in-time, in-place, no-device assessments.

I've deliberately left this one until last, and ironically, it's where many institutions went first when ChatGPT was released. But this doesn't necessarily mean examination-style assessments.

Group work, orals, seminars, practicals, simulations, vivas, brainstorming with post-it notes, debates, marker pens on butcher's paper... There are plenty of methods that predate GenAI by a few centuries and still work.

Advantages:

- Easy to monitor and secure; with no access to devices and no way to do what the Victorian police call “sneaky face” (looking at a phone while driving or, in this case, under the desk), there's no GenAI to worry about here. We might call these ‘Level 1’ assessments in our AI Assessment Scale.
- Relevant, engaging, and authentic; these assessments can be modeled on real-world and authentic experiences, such as carrying out a practical task or a simulation.

Disadvantages:

- Unfortunately, this type of assessment is hard to scale. It might work well for tutor groups or K-12 classes, but it becomes unwieldy in a cohort of 100+ students.
- No online mode. Short of relying on lockdown browsers and creepy surveillance tech, there's no way to guarantee “no devices” in an online setting. I'll be writing more about GenAI and online teaching at a later stage because it's a whole different ballgame.

So, those are six suggestions for assessments that account for GenAI but don't rely on detection tools. None of them are perfect, and each comes with its own set of

challenges, but I believe they're a step in the right direction as we navigate this new landscape of education in the age of artificial intelligence.

GenAI Strategy: Attack Your Assessments

Published: 2024-05-13 | Original: <https://leonfurze.com/2024/05/13/genai-strategy-attack-your-assessments/>

This post is part of a series exploring how faculty and middle leaders can begin to think strategically about Generative AI. In the previous post, I gave an overview of a six-step strategy, beginning with establishing a clear vision. In this post, I'm going to go deeper into attacking your assessments.

UPDATE: Less than 24 hours after posting this, OpenAI announced the release of GPT-4o, their new FREE model. This makes the paid model, GPT-4, available to all users via web, mobile, and a new desktop macOS app. GPT-4o includes image recognition, file upload, image generation, and advanced language, code, and reasoning capabilities. When I published this article, I recommended using a paid license to attack your assessments. That is no longer necessary. Go to chatgpt.com and try your assessments in GPT-4o as soon as you can.

I'm going to be blunt (and if you've seen me speak recently, this will not surprise you). It doesn't matter if you aren't interested in AI. It doesn't matter if you personally don't use AI. It certainly doesn't matter if you believe AI won't impact your discipline. Because the brutal reality is that Artificial Intelligence, including the current GenAI flavour of the month, will impact every industry that involves digital technologies: and that's most of them.

In the last post, I showed a few snippets of GenAI completing a variety of tasks, including senior secondary certificate level Maths, Physics, Biology, and Art. And in Step 2 of the strategic planning, I am encouraging you to do the same: attack your assessments. It is the only way of demonstrating to yourself and your colleagues that *it really doesn't matter what you think* about Generative AI.

It's a cold, brutal reality. As an English teacher, I have to ask to hard questions about the subject that I've taught for over fifteen years, and the methods I have used to

teach. Where I've encouraged students to use digital technologies for research, writing, or creating multimodal texts, I now have to wonder whether those skills will be supplanted or even lost as a result of AI. I have to ask myself whether it might be better to go back to "no tech" methods of instruction.

And in some cases, "no tech" will surely be the answer. I don't think I could confidently teach students the fundamental skills of writing analytical essays if they have access to a phone, laptop, or other device. I have *absolutely no way of knowing* whether the work I'm seeing at the end is the student's, or has been created partially or entirely with GenAI. So I have two things I need to do right now:

1. Decide which skills are fundamental to my discipline, and which absolutely need to be learned slowly, methodically, and without offloading onto technology.
2. Decide which skills and content I *can* (or must) offload, knowing that GenAI is now competent across a broad range of multimodal skillsets.

This is where it becomes necessary to *attack your assessments*. Whatever you are teaching, and whether you're in K-12 or Higher Education, your assessment tasks are more vulnerable than you think to Generative AI.

You can access a free ebook on the AI Assessment Scale with over 50 activities for the 5 levels by signing up for the mailing list here:

Attack your assessments

I stepped out this activity [in the first post in the series](#), and I'll outline it again here. I think you need a systematic approach to attacking your assessments and it will take a little investment of time and money (for a quality paid model) to get this right.

Here are the step-by-step instructions for the faculty meeting activity:

1. **Before your next faculty meeting:** Ask staff members to share one or two of their current assessment tasks by either dropping them into a shared folder or sharing the link to the assessment in the school's learning management system. Identify a couple of faculty members who are confident using a range of generative AI tools, and make sure they have access to either ChatGPT Plus or Claude 3 Opus (both \$20 USD/month), or Midjourney for image generation (\$11

USD/month). If no one in your faculty is confident with AI, consider adopting the role yourself.

2. **Brief the chosen GenAI-using faculty members:** Instruct them to adopt the role of a student deliberately using GenAI to complete as much of the shared assessment tasks as possible, as quickly as possible. Encourage them to be creative in their approach, depending on the subject area. For example:
 - In visual arts, they could use image generation platforms to create amateurish photography portfolios, using language prompts to avoid telltale signs of AI.
 - In maths and science, they could investigate ChatGPT Plus's code interpreter or GPT-4's mathematical reasoning abilities to solve problems directly or write and execute code.
3. **During the faculty meeting:** Share the AI-generated examples with your colleagues, which may include essays, images, audio, or video. For maths and science teachers, demonstrate how AI can be used to solve problems or write and execute code. Use a real-world example, such as the questions from the VCE or HSC exams, or your first year university assessments, to show how AI can solve the problem and compare its answer to the examiner's report or criteria.
4. **Facilitate a discussion:** Encourage faculty members to share their thoughts and observations about the AI-generated content – there will be push back, possibly driven by fear. Discuss the vulnerabilities in the current assessment tasks and how AI could be misused by students. Address any “this doesn't affect me” attitudes by highlighting the rapid development of AI technologies across various domains.
5. **Brainstorm solutions:** Divide the faculty into small groups to brainstorm ways to modify or create new assessment tasks that are more resistant to AI misuse. Encourage the groups to consider alternative assessment methods, such as project-based learning, collaborative work, or in-person demonstrations of skills.

Examples from across the disciplines

I shared a few of these in the previous post, and I'll include some new ones here. I want to clearly demonstrate how vulnerable different domains are to GenAI. Importantly, this is not intended to reinforce “pen and paper” methods of

assessment. Rather, my intention is to prove that GenAI can handle certain knowledge tasks across domains, and we therefore need to update our assessments to make use of the technology where appropriate. We also need to really clearly identify which skills we are assessing and *why*, for any given assessment.

Of course, all of these examples are from Australian senior secondary studies. It's worth seeing how GPT-4 compares at more advanced problems. So here's GPT-4 attempting a problem from MIT's Open Course *Principles of Chemical Science*, Unit II: Chemical Bonding & Structure.

As you can see in the following video, it handles the task well. The first question is misinterpreted slightly, but parts B and C answered correctly (part C differs, but only due to a rounding error). The most important aspect here is that GPT-4 is able to use its Code Interpreter feature to handle the more complex mathematics. It isn't relying on the predictive model, but instead delegating the task to the programming language Python, which is more than capable. Of course, since these materials are open access they could be present in the dataset. But without many duplicate copies, it's unlikely the model has "learned" the correct answers.

Confronting reality

The brutal reality is that Generative AI is here to stay, and it's going to impact every aspect of education, whether we like it or not. As educators, we have a responsibility to confront this challenge head-on, and that means attacking our assessments to identify vulnerabilities and explore alternative methods that can withstand the AI onslaught.

It doesn't matter if you're teaching English, maths, science, or art – GenAI is coming for your assessments. We've seen examples of AI models like GPT-4 acing senior secondary certificate exams and even tackling complex problems from MIT's Open Course Principles of Chemical Science. So, what do we do? We need to get systematic. We need to bring our faculty together, share our assessment tasks, and let our colleagues loose on them with sophisticated GenAI tools. Let them be creative, let them push the assessments, and let them show us just how vulnerable our current methods really are.

ChatGPT using GPT-4o (the new free model, see update at the top of the article) on the VCE Specialist Mathematics examination. I subsequently

tried both papers, and it scored 100%. [Go to my LinkedIn](#) and scroll back through the activity to see more of these experiments with the new free model.

Once we've seen the reality of the situation, we need to have an honest discussion about which skills are fundamental to our disciplines and which ones we can (or must) offload to AI. We need to brainstorm new assessment methods that are more resistant to AI misuse, whether that means project-based learning, collaborative work, or in-person demonstrations of skills.

The bottom line is this: we can't ignore Generative AI, and we can't pretend it won't affect us. We need to embrace the change, adapt our teaching and assessment practices, and find innovative ways to leverage AI to better prepare our students for the challenges and opportunities of the future.

Maybe it's a brutal reality, but it's one we have to face. So, in your next faculty meeting, get ready to attack your assessments.

GenAI Strategy: Bullets then Cannonballs

Published: 2024-05-17 | Original: <https://leonfurze.com/2024/05/17/genai-strategy-bullets-then-cannonballs/>

This post is part of a series exploring how faculty leaders can develop generative artificial intelligence strategies specific to disciplines in K-12 and higher education. The first post gave an overview of the strategic planning process. The next post followed up by [encouraging you to attack your assessments](#). If you haven't already, you should read those posts before progressing on to this stage of the strategic planning.

The introduction of mainstream generative artificial intelligence at the end of 2022 has left the education sector, from kindergarten through to higher education, in a state of flux. For faculty leaders and executives in education providers, this means that we should be developing strategy which can withstand the next few years.

Of course, we've seen in the last 18 months how hard it is to predict the trajectories of these technologies. Nobody, including OpenAI, knew how much of an impact GPT would have on every sector, including education. Nobody could have anticipated Microsoft going all in on generative artificial intelligence or the speed at which it would be possible for developers to turn out multimodal applications like the upcoming [Project Astra from Google](#), or [OpenAI's recently released GPT-4o](#).

But just because the pace of change is rapid doesn't mean we can only plan month by month or day by day. We should still aim to strengthen parts of the education system that need to stand the test of time.

There are, of course, outdated parts of this system which generative artificial intelligence (and before that, remote learning) exposed. In the previous post, I encouraged you to sit down with your faculty and [attack your assessments](#), using a sophisticated model such as GPT-4o or Claude 3 Opus to tackle examinations and assessment tasks in your discipline. If you completed that exercise, you've probably seen that these models absolutely have the capacity to address previously human-

only domains such as logic, mathematical reasoning, and finding meaning from images.

These applications are still far from perfect, but the technology is very capable; it's time to do something about it.

Spaceships, Bullets, and Cannonballs

Whenever a new technology emerges, there's a temptation to rush right in and start building on top of it. We've seen this already in education with generative artificial intelligence, largely in the form of educational chatbots and tutors which use large language models to create interactive assessments that guide students through curriculum material.

Organisations like Khan Academy have partnered with OpenAI, and we know that millions, if not billions, of dollars have already been spent on fine-tuning these tutor bots. In education, we've seen smaller scale variants of this, from departments of education like New South Wales and South Australia developing chatbots in league with Microsoft, to individual schools, universities, and institutions creating chatbots with their internal data.

But I think that we should scale back our approach before worrying about system-wide deployments of generative AI. Of course, the technology moves quickly, and it may feel like there is an imperative to build now. [In fact, I've written about this before, reflecting on a problem called the wait calculation.](#)

Imagine an interstellar voyage. You can build spaceships with the technologies that we have now that might take, say, 500 years to reach the destination. Or you could wait for 50 years and, with technology advancements, build a ship that will get you there in 100, overtaking the original 2024 spaceship while it's still chugging through the first part of its journey. Sometimes it is genuinely better to wait and see. Collins's "fire bullets then cannonballs" analogy speaks to a similar logic.

Imagine ships approaching each other through a dense fog. The captains give the order to fire without being able to see their opponents; they could be shooting at nothing at all.

One captain loads up the cannons, filling them with gunpowder and launching cannonball after cannonball into the ocean. The other captain instructs his crew to

load their rifles and fire a barrage of shots in all directions until they hear the telltale clink of bullets on timber. The moment they've identified that the shots are landing, he switches to the ship's cannons.

You can put all of your time and energy into building big, flashy systems that may or may not work. Or you could spend the next few months trialling smaller experiments, and when you have evidence that your approaches are landing, and *if and when* you have evidence that your approaches have a positive educational outcome, *then* you can start to fire cannonballs.



I'm making no excuses for mixing up my interstellar travel and naval battle analogies, although I will say that strategic planning is overladen with analogies from war. But this *is* serious stuff; lives are genuinely at stake. If billions of dollars worldwide are invested into technologies which have been rushed into, deployed too early, and ultimately fall flat, then students will suffer.

If chatbots are rolled out into schools with assurances of safety and guardrails that remove bias, and then these systems are proven to be insecure or faulty, students will suffer.

If schools and universities run headlong into making decisions based on FOMO rather than evidence, then *students will suffer*.

Which bullets will you fire first?

Having familiarised yourself with local and national policies around generative AI in [step one](#), then [taking on your assessments in step two](#), you'll now need to decide which small experiments you're going to carry out on a faculty level that you can measure and get rapid feedback on. If you teach in design and technology, STEM, engineering, or entrepreneurial education, then you've got an advantage here, because I'm going to suggest that you run through a design thinking cycle in your next faculty meeting to determine which bullets you'll fire first.

You might only have an hour in your faculty meeting to go through this process, so I'm also going to suggest that you use generative AI extensively in all stages, but that you always fall back on the expertise of the people in the room. At a very high level, you are going to decide whether your first experiments will be for teachers (the educators), the students, or both.

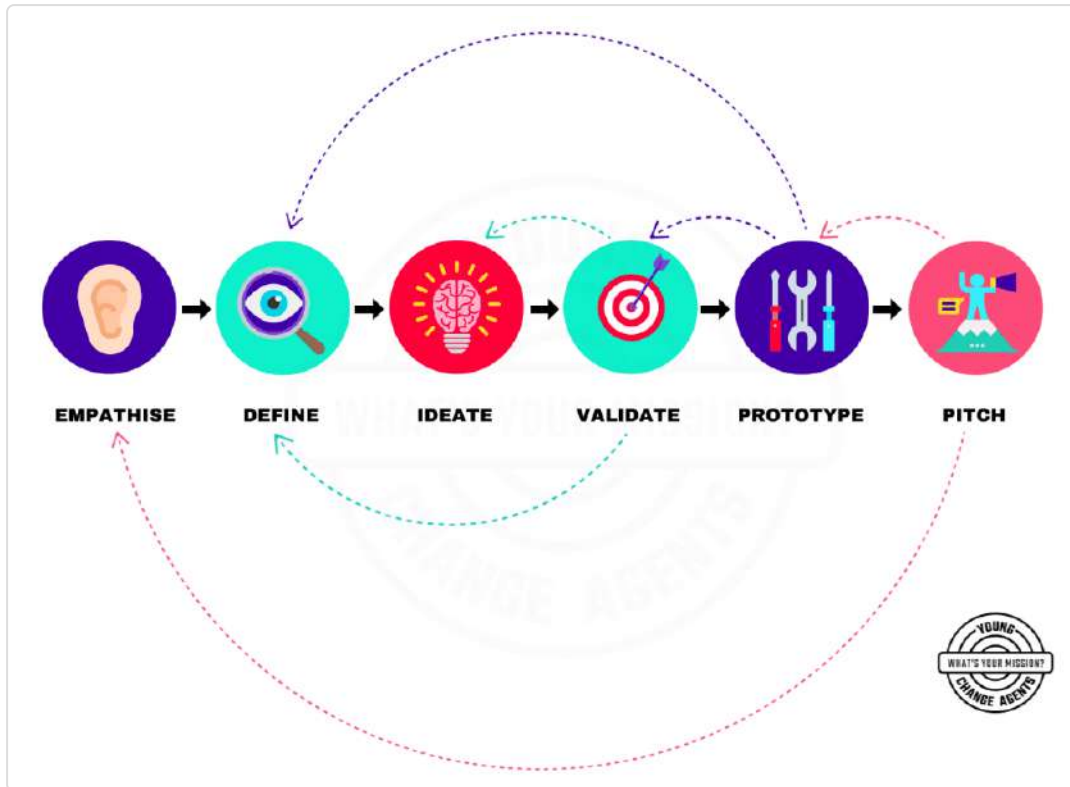
To an extent, this will be determined by the level of familiarity with the technology in your faculty. If everyone has little to no experience with generative AI, it will be much more productive to try some small experiments where teachers or educators are using the tools for themselves in their day-to-day work before worrying about using it in the classroom. If, on the other hand, you've been experimenting with these technologies for a while, it may be time to introduce it to students. Whichever approach you adopt, you will need to create a process leading to measurable outcomes.

Here are the steps you're following:

1. Quick recap on everything that you've learned so far, from the initial documentation review, draft vision, and attacking your assessments.
2. Clarify your audience
3. Design Thinking workshop

You can access a free ebook on the AIAS with over 50 activities for the 5 levels by signing up for the mailing list [here](#):

The Design Thinking Cycle



<https://youngchangeagents.com/educators/design-thinking>

Empathise: Begin by identifying the audience for this experiment. Is it yourselves, the educators? Is it the students? Or is it both?

What can you do right now to gather evidence of what this audience needs? If you're focusing on yourselves, you should know most of this already from the previous stages. If you're a bit further along and have decided to work with students, this would be a good opportunity to invite some of them into the discussion to hear their opinions on generative AI and what they think might be the advantages and disadvantages of using generative AI in your discipline. If you don't have time to speak directly to students at this stage, don't worry, because you will do that later. You might choose instead to have the discussion amongst yourselves or to use a chatbot with a prompt like the following:

You are a [e.g., secondary school, university] student studying [discipline]. We are the faculty team and have discussed generative artificial intelligence and its implications for our subject. We want you to identify some of the advantages or disadvantages of using

generative artificial intelligence in this discipline from your perspective.

If you're in a small faculty or if people are insecure about generative AI, you could, of course, tweak this prompt to have the chatbot join in as a faculty member rather than as a student. Neither of these replaces human feedback.

Define: Next, define the exact problem you're trying to solve. For example, is it increasing staff understanding? Understanding how students interact with chatbots?

Since you're a group of educators, consider framing this with the language you might use for an objective, such as "to understand" or "to know".

If you really must, this is an opportunity to get out post-it notes. I personally have a "no post-it" policy when it comes to strategy (and I think I'm allergic to butcher's paper), but you do whatever works. Try to get as close to the real problem as possible, and remember, small experiments not huge commitments.

Ideate and validate: You're going to brainstorm as a group for small experiments.

At this stage, it might be good to capture the discussion as an audio recording or take minutes of all of the ideas. If you do use the forbidden butcher's paper and post-it notes, encourage people to write legibly. Take photographs of those materials to use with an AI model with image recognition like Claude, Copilot, Gemini, or GPT-4o.

Brainstorm for as long as you have the time to generate as many ideas as possible. Consider using a structured approach to the brainstorming, such as a visible thinking routine. This can help to remove some of the fear of the blank page and encourage more meaningful ideas. Use the same processes that you would use with students in your classroom. There's nothing worse than when I work with strategic planning in education and I see people using brainstorming activities that we *know* wouldn't work with students (list as many words as you can that are associated with this school...). You can't just ask people to pull ideas out of thin air; idea generation often benefits from structure.

Now, I want you to refine those ideas using your generative AI platform of choice. I would recommend Claude or GPT-4o for this stage because they handle a lot of data and do an impressive job of interpreting language and images. If you recorded the discussion, use an application like Otter.ai to transcribe it and upload the transcript

to Claude via copy-paste or by uploading the Word document or PDF. Provide the following prompt:

We are a faculty team trying to create some small experiments with generative artificial intelligence that would work in our discipline. These experiments will be carried out with teachers/students [in the subject area]. To help us clarify and extend these ideas, we would like you to carry out the Generate-Sort-Connect-Extend visible thinking routine. <copy/paste initial ideas>

Prototype: Once you have a refined set of ideas from the AI-assisted brainstorming session, it's time to select a few of the most promising ones to prototype.

These should be small, manageable experiments that you can implement relatively quickly. For example, if you're in a math department, you might prototype using ChatGPT to generate practice problems for students. Or if you teach writing, you could experiment with using AI to provide initial, low stakes feedback on student drafts. The key is to keep the prototypes small and focused.

Test: With your prototypes ready, the next step is to test them out. This is where you gather data on how well they work in practice. Set clear metrics for success upfront. Are you looking at student engagement, quality of work, time saved for teachers, or something else? Collect both quantitative and qualitative data – survey students and teachers, look at student work, track time spent. Document what works and what doesn't.

Here's a suggestion for a feedback survey to use with your faculty:

1. Which AI experiment(s) were you involved with? [Short answer]
2. On a scale of 1-5, how effective do you feel the AI was in enhancing the educational experience? (1 = not at all effective, 5 = extremely effective)
3. What worked well about the use of AI in this experiment? [Long answer]
4. What challenges or concerns did you encounter with the AI? [Long answer]
5. Did you notice any positive or negative impacts on student learning? Please explain. [Long answer]

6. How do you think the use of AI in this experiment could be improved? [Long answer]
7. On a scale of 1-5, how comfortable are you with the idea of integrating this AI application into your regular teaching practice? (1 = very uncomfortable, 5 = very comfortable)
8. What support or resources would you need to effectively use this AI tool in your teaching? [Long answer]
9. Do you have any concerns about the ethical implications of using this AI in education? [Long answer]
10. Any other comments or feedback you'd like to share? [Long answer]

Importantly, start to get feedback from *all* stakeholders. What do students think of the AI-powered tools? How do teachers feel it impacts their work? Do parents or administrators have concerns? Testing isn't just about validating your ideas, it's about learning and iterating. We'll do more of this in the next stages of the strategic planning when we start to communicate the strategy, evaluate and review.

GenAI Strategy: Update your Assessments

Published: 2024-05-21 | Original: <https://leonfurze.com/2024/05/21/genai-strategy-update-your-assessments/>

This post is part of a series on developing generative AI strategy from faculty leaders and teaching and learning teams. Having been both a faculty leader and a Director of Teaching and Learning, I understand that a great many school strategies get lost in translation when they reach middle leadership and have to be put into action.

Because artificial intelligence will impact every discipline, I think we need to lead some of this change from the middle. If you haven't already, check out the [previous post](#) in this series, first, outlining the whole strategy, then [attacking your assessments](#) and carrying out some [small experiments](#).

Assuming you've followed the steps and [attacked your assessments](#), it's now time to look at the assessments you already have and find ways to update them to account for this technology. I don't want faculty leaders and educators to completely redesign all of their curricula and assessments. Some tasks will need significant revisions and others only small tweaks.

If you are not familiar with the Artificial Intelligence Assessment Scale, developed by Dr. Mike Perkins, Dr. Jasper Roe, Associate Professor Jason MacVaugh, and myself, then I recommend you familiarise yourself with the following resources before going any further with this article:

- [Assessment Scale Version 2 Blog Post](#)
- [AI Assessment Scale Pilot Project Blog Post](#)
- [AI Assessment Scale eBook](#)
- [JUTLP Volume 21 Number 6: AI Assessment Scale Pilot](#)
- [The AI Assessment Scale \(AIAS\) in action: A pilot implementation of GenAI supported assessment \(preprint\)](#)

You can access a free ebook on the AIAS with over 50 activities for the 5 levels by signing up for the mailing list here:

This comprehensive set of resources explains the rationale for the AI Assessment Scale and why such a scale is necessary. It includes lots of examples across disciplines which apply to both K-12 and higher education.

The AI Assessment Scale in Brief

We developed the AI Assessment Scale because we acknowledged that schools and universities needed something more nuanced than a binary “use/don’t use” approach to this technology.

Students have also told us through [various studies](#) that they want education providers to help them understand how to use generative AI technologies responsibly and ethically. Here in Australia, the national [Framework for Generative AI in Schools](#) includes two explicit guiding statements which speak to the need for a structured approach to using artificial intelligence:

In Core Principle 1: Teaching and Learning, we find the following:

- **1.5 Learning design:** work designed for students, including assessments, clearly outlines how generative AI tools should or should not be used and allows for a clear and unbiased evaluation of student ability.
- **1.6 Academic integrity:** students are supported to use generative AI tools ethically in their schoolwork, including by ensuring appropriate attribution.

These tools are not to be banned, and students must be supported in their use, but assessment tasks need to be absolutely clear on how AI can and cannot be used in a given situation.

The AI Assessment Scale runs from Level 1 to Level 5, ranging from No AI to Full AI. Our most recent version, published in the *Journal of University Teaching and Learning Practice*, is our most up-to-date version.

1	NO AI	<p>The assessment is completed entirely without AI assistance. This level ensures that students rely solely on their knowledge, understanding, and skills.</p> <p>AI must not be used at any point during the assessment.</p>
2	AI-ASSISTED IDEA GENERATION AND STRUCTURING	<p>AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.</p> <p>No AI content is allowed in the final submission.</p>
3	AI-ASSISTED EDITING	<p>AI can be used to make improvements to the clarity or quality of student created work to improve the final output, but no new content can be created using AI.</p> <p>AI can be used, but your original work with no AI content must be provided in an appendix.</p>
4	AI TASK COMPLETION, HUMAN EVALUATION	<p>AI is used to complete certain elements of the task, with students providing discussion or commentary on the AI-generated content. This level requires critical engagement with AI generated content and evaluating its output.</p> <p>You will use AI to complete specified tasks in your assessment. Any AI created content must be cited.</p>
5	FULL AI	<p>AI should be used as a 'co-pilot' in order to meet the requirements of the assessment, allowing for a collaborative approach with AI and enhancing creativity.</p> <p>You may use AI throughout your assessment to support your own work and do not have to specify which content is AI generated.</p>

<https://open-publishing.org/journals/index.php/jutlp/article/view/810>

Auditing Your Assessments

Rather than starting from scratch with all of your assessment tasks, or panicking that generative AI means you'll have to throw everything out, I suggest you carry out an audit of your major summative tasks to see if they already align with one of the five levels. To help with this process, we'll be developing an AI assessment audit tool which you can use in a faculty meeting or distribute to your team members and have them audit the assessment tasks that they are responsible for.

Until that tool is available, you can carry out the process yourself by placing existing assessments at the most appropriate level and discussing how the task might be adjusted to bring it up or down. Avoid getting bogged down – limit this to the really crucial graded assessments and end-of-unit assessments. Or, if you have project-

based learning, align them with the overall grade and final criteria for the whole project.

It may be useful to have some double-up if you have multiple faculty teaching the same assessments. And then do this individually rather than as a group for a diverse range of perspectives – the more ideas you can collect, the more robust your final judgment will be. Here are some questions you might ask:

- How much does it matter if the student remembers or memorises content?
- Is it an examination? Does it need to be?
- How much does it matter if the student is demonstrating their ability to write or their literacy skills?
- How much does it matter if the task involves creativity or critical thinking?

None of this is set in stone. You may find cases where your assessments don't neatly fit into one particular area of the scale. We designed the scale to be flexible from day one and published it open access so that you can take the idea of the scale and adapt it to your own context.

After auditing your assessment tasks, [you may decide to create a version of the scale which is more suitable for your organisation or faculty](#). For consistency's sake, I would recommend having an institutional version of the scale which allows some variations in the wording for different faculties but keeps the integrity of the levels.

After the Audit

After staff have completed the audit, I would recommend the faculty leader collates this information and then presents something back during a meeting. You may find some inconsistencies between individual staff members' perceptions of assessment tasks. You might find that certain staff wish to do everything at the "No AI" level because of concerns that students will cheat or misuse the technology. These are valid concerns, and you as the faculty leader needs to manage any of those conflicts.

In the spirit of the previous step, "Bullets then Cannonballs," I would suggest that you pilot the AI Assessment Scale only for the assessments in one term, or focus on a particular subject or year level before attempting to apply the AI Assessment Scale to all of your assessments.

In the next step, we will look at evaluating and communicating what you have learned with colleagues and leadership.

GenAI Strategy: Evaluate and Communicate

Published: 2024-05-27 | Original: <https://leonfurze.com/2024/05/27/genai-strategy-evaluating-and-communicating/>

This is the fifth in a series on generative AI strategy for faculty leaders and teaching and learning teams. If you've been following along, you will have:

1. Prepared yourself with local and national school guidelines on AI
2. Attacked your assessments
3. Tried some small experiments
4. Started updating your major assessments with the AI assessment scale

You should by now have collected enough data from staff and students to begin making some judgments and to communicate those findings with your colleagues and leadership.

One of the biggest unknowns of generative artificial intelligence in education (or even more broadly, technology as a whole) is whether or not it actually has a positive impact on learning. Unfortunately, there are [decades of studies which suggest that as far as edtech is concerned, the opposite is true.](#)

AI might have a neutral or even negative impact on learning due to the risks posed by digital technologies and devices on student's attention spans. Anyone who has been in a classroom knows how easy it is for students (and ourselves) to become distracted. Despite all of the hype and rhetoric around AI revolutionising education and the glossy sales pitches from companies like OpenAI and Google, I will not commit to saying that AI will have a positive impact on learning until I see some evidence.

I think it's vitally important that educators on the ground carry out these small experiments and evaluate the efficacy of generative AI for themselves, so that we're not simply being told by technology companies that AI will "revolutionise education".

You can access a free ebook on the AIAS with over 50 activities for the 5 levels by signing up for the mailing list here:

Evaluate Your Findings

In the previous steps in this strategic planning, you should have carried out some [small experiments](#) (firing bullets, then cannonballs) and began to address your major assessments, perhaps aligning them with the AI Assessment Scale. I encouraged you to collect feedback from students and/or teachers, depending on the kinds of experiments you carried out. And it's now time to evaluate all of that feedback.

Since this is not a peer-reviewed study, we can perhaps be a bit fast and loose with the methods and use generative artificial intelligence to speed the process along. If, like I suggested in your assessments, you subscribed to Claude from Anthropic, this would be my recommended tool for going through your data. However, by the time you're reading this, you may well have access to GPT-4o for free, in which case, you can certainly try that.

The intention is to gather all of your data in one place, and then crunch it with AI. If you sent out student surveys or teacher evaluations, put all the open-ended question notes into a single CSV or Excel file. (If you used Microsoft Forms or Google Forms, you could also export them as PDFs and merge them together but de-identify the results first) Because Claude can take a large amount of text in its context window, there shouldn't be any issues even if you have sampled a large group for feedback.

If you followed my advice in the previous posts, your feedback will likely have long-form responses from staff or students. And we're going to use the AI to synthesise those responses. For the numerical responses, you can obviously just use the built-in graphs in Microsoft or Google Forms.

In the video below, I'll demonstrate the process. I'm using feedback from participants in one of my courses because obviously, I'm not evaluating my faculty's use of AI. And importantly, I have de-identified the results before using them with AI, deleting the names and email addresses collected during the survey. I've demonstrated this in both GPT-4o (MacOS desktop version) and Claude for comparison:

Watch the video and follow the process with your own data. If you have collected information from students and teachers, then do it separately for each. **Remember**

to de-identify all of your data before using any GenAI platform. Prepare this evaluation for your next faculty meeting for discussion.

Discuss Your Findings

Make no assumptions about whether your findings will be positive or negative. As I said earlier, I can't absolutely promote the use of AI in education is beneficial until I've seen that it has a positive impact on things like student learning or teacher workload. Everything up until that point is just conjecture and my own opinion.

In the spirit of [confronting the brutal reality](#), discussed in the first post in this series, don't try to gloss or tidy up the results. If your results are mostly negative (for example, teachers talking about the learning curve being too steep, not having time to use or learn to use the technologies, or finding them a distraction in the classroom), then address those issues up front. If student feedback indicates that they don't enjoy interacting with chatbots, that they feel they are not learning, or it raises any other issues, confront these as well. Do so with an open mind. Ask questions like the following:

- Did the way we conducted these experiments contribute to the negative outcomes?
- Have our perceptions or discussions of these technologies coloured the output?
- Is my personal bias as a faculty leader contributing to these results?
- Are these results due to technological or pedagogical issues?
- Could we try another experiment? Might it have different results?

If you get overwhelmingly positive results, I would still encourage you to ask the same kinds of questions. Reflecting on your processes whilst carrying out your small experiments and gathering that data is an important step before we move on to communicating your findings.

Whether your responses are positive or negative, it's worth acknowledging that the broader encroachment of these technologies into education will not be slowed. You might unearth some serious problems with the ways that students are interacting with chatbots, but that will not stop them from interacting with chatbots. The next step is to clarify and update your earlier draft vision and decide how you will communicate.

Clarify Your Vision

In the first step, you set out to draft a vision for the faculty based on school, local, national, and international guidance on incorporating generative AI. Now, having been through the process of testing it out in your discipline with your staff and students, you should be able to refine that. Perhaps you found that despite your best efforts, using generative AI really didn't improve the quality of students' work or their understanding. Or maybe you found that encouraging teachers to use generative AI reduced their workload significantly. Incorporate these findings into your original vision. For example, if your vision started off as something like this:

For students in our faculty to ethically and appropriately use generative AI tools

Sharpen that vision up to something like this:

Guide students through ethical and appropriate use of generative AI technologies whilst prioritising technology-free instruction for fundamental skills

Or if your vision began like this:

To explore ways that generative artificial intelligence can reduce teacher workload in our faculty

You might make it more specific by incorporating some of your findings, like this:

Incorporate ethical and appropriate automations into planning and communication processes for all staff in our faculty to reduce workload

Once you have your updated vision, it's time to pull everything together and get ready to communicate your strategy.

Communicate

As the faculty leader, you're now going to develop a generative AI strategy one-pager. This needs to be simple and succinct because this might be communicated to school leadership (including assistant principals and principals) who have not been part of this strategic planning process. Or you may be communicating it to other faculties who have not yet been through the process themselves. You might also be

communicating it to parents or external stakeholders like the school or university board.

Again, you can use generative AI to help out at this stage. But I would encourage you to put all of the materials on the table for thorough discussion with your human colleagues before handing the synthesis over to the AI. Here's how I would suggest you write your one-pager:

1. Gather up all of your materials, starting with the [initial documentation that you reviewed](#) (international, local, and school guidelines) and your draft vision. Quickly revisit these ideas and see if there's anything you've missed, or drifted away from. For example, if you're in an Australian K-12 context, are you still adhering to the core principles of the Australian Framework?
2. Gather your reflections from attacking your assessments and updating your assessments. Write a short statement describing how you identified vulnerabilities in your existing assessments and how you have experimented with updating your assessments accordingly.
3. Write a brief statement about your small experiments, whether they were for staff or students. Try to summarise your findings as concisely as possible in a paragraph or dot points.
4. Write out your updated vision and make sure everybody is comfortable with it. (Although please, for the sake of everyone in the room, don't spend an hour debating synonyms.)
5. Synthesise your observations about assessment and your findings from your small experiments into a set of guiding principles or strategic directions which support your vision.
6. Create a one-pager with the vision, the strategic directions, and (if you feel it necessary) a brief explanation of the process by which you reached these conclusions. See the example below.
7. Last of all, communicate this with your colleagues, community, and students. You'll probably go up the line first to the executive team to get their approval, and then perhaps communicate horizontally to other faculties, and finally to students, parents, and the broader community.

I've run parent information sessions with dozens of schools and thousands of parents, and I can safely say that there is an incredibly high level of interest from

parents who want to know if schools are addressing these technologies and their implications for teaching and learning. By going through this faculty strategic planning process, you've demonstrated your commitment to doing so.

Example English Faculty GenAI Strategy One-Pager

English Faculty AI Vision and Strategy

Vision: To understand the potential of generative AI to enhance English teaching and learning in ways that are safe, ethical, inclusive and aligned with the Australian Framework for Generative AI in Schools.

Key Principles and Directions:

1. Educate students at appropriate levels

- Educate students about how generative AI works, its capabilities and limitations
- Prioritise the instruction of critical thinking to evaluate AI-generated content and identify potential biases
- Integrate AI tools into the English curriculum in age-appropriate ways, guided by national regulations

1. Use AI to support teacher expertise

- Explore using AI to assist with curriculum development and differentiated instruction
- Maintain teacher control over instructional decisions and student evaluation, particularly for high stakes assessments
- Provide ongoing, subject specific training and support for staff on effective AI use in English

1. Adapt assessments for generative AI

- Review all existing assessments and identify ways students could misuse AI
- Experiment with varied assessment types that measure higher-order skills to which AI does not apply

- Clearly communicate AI use policies in assessments and evaluate the impact of AI use through ongoing measurement of student data

1. *Ensure accessibility, fairness and student wellbeing*

- Provide equitable student access to AI technologies used in English
- Monitor AI use and impacts to prevent harm, discrimination or privacy breaches
- Educate students on responsible AI use, including copyright and attribution

1. *Open experimentation, evaluation and dialogue*

- Continue to conduct small trials of AI in English to guide faculty understanding
- Actively monitor developments in generative AI and iterate on our approach, whether that means increasing or decreasing AI use
- Engage staff, students and parents to communicate vision and incorporate feedback

This strategy was informed by an extensive review of AI in education guidelines, consultation with staff and students, and trials of AI technologies in English classes. While still a work in progress, it provides an initial roadmap for putting our AI vision into practice in a thoughtful, measured way. We will continue refining our approach through ongoing experimentation, evaluation and stakeholder engagement.

Reflect and Review

This is the final stage of the strategic planning, but it doesn't really warrant an entire separate article because you've already done most of the hard work. You now need to set some regular checkpoints for reviewing the strategy and continually evaluating its effect. The moment you identify that something is not having a positive impact on teaching and learning, stop doing it. There are more than enough ineffective processes in the education system without us pouring AI special sauce over the top of them.

I would encourage you to continue in the spirit of experimentation, occasionally bringing in new small experiments (firing a few more bullets). And when you identify

your targets, whether it's student engagement, learning, teacher workload, or something else, lean in and see if you can sharpen your faculty strategy.

Most importantly, share what you are learning with your colleagues and the community. Share it with me through the contact form on this website, by email, or on LinkedIn. Nobody in education knows the answer to whether these technologies will have positive, neutral, negative, or mixed impacts on education. The more of us that carry out these small experiments and share what we find, the better off everyone will be.

Over the coming weeks, I'll be turning this series into an online course to guide faculty leaders through the process. Make sure you join the mailing list to stay up to date on those announcements.

GenAI Strategy for Faculty Leaders

Published: 2024-05-01 | Original: <https://leonfurze.com/2024/05/01/genai-strategy-for-faculty-leaders/>

When I work with K-12 schools, the approach to generative artificial intelligence (AI) guidelines generally begins with an audit of current policy documents that relate to digital technologies, such as user agreements, cyber safety policies, and communications. The goal is to identify areas where GenAI needs to be taken into account, without necessarily writing entirely new policies.

Once a broad strokes approach is established, it usually falls back to the assistant principals and directors, particularly people with a role in the school like:

- Deputy Principal
- Assistant Principal Teaching and Learning
- Director of Innovation
- Director or AP of Student Wellbeing
- Head of Faculty
- Year Level Coordinator

While it's absolutely necessary to have some high-level policies and guidelines, the appropriate adoption of Generative AI is being driven by middle leaders to translate to classroom practice.

Generative Artificial Intelligence in Different Subject Areas

I've written before about the importance of faculty-level strategy. In this short series of posts, I'll be talking about strategic planning for faculties handling artificial intelligence. Every subject will be affected by artificial intelligence differently; what is successful in English will be entirely different in mathematics, design technology, or music. There are also the ethical and practical considerations of the technology that students might come across, which vary from topic to topic.

Of course, there are some transferable skills which might be helpful for faculties to deliberately take on, distributing the load of teaching students how to use generative AI. For example:

- The English faculty may focus on the use of generative AI writing tools such as large language models and applications like ChatGPT.
- The arts faculty may be better served focusing on the ethical and creative implications of image, video, and audio generation.
- Subjects like science, mathematics, and digital technologies may explore the use of generative AI for coding, data science, and data analysis.
- Foreign language subjects might look at multilingual translation tools.
- Health and Physical Education classes may look at the impact of algorithms in healthcare or how AI might be used for medical assistants, personal training, and coaching.
- Design and technologies teachers might incorporate generative AI into various stages of thinking, from the use of large language models for idea generation to next-generation 3D modelling for rapid prototyping.

The applications of multimodal GenAI are vast, and in many fields, we've barely scratched the surface of the potential for AI to support the full range of subjects. Until we develop methods for subject experts to grapple with these technologies in their own disciplines, we will stay at the surface level.

This first post lays out a general strategic framework that any faculty leader can adapt to bring discussions of this technology into their domain. In subsequent posts, I will be inviting faculty leaders from across a range of disciplines to contribute ideas on how AI can be used in their disciplines.

Step 1: Review Frameworks and Policies and Establish a Faculty Vision

The first step in your faculty strategic planning process should be to get the lay of the land. If you're not familiar with key local, national, or international frameworks on artificial intelligence, you should spend some time reviewing these documents beforehand. This should not occupy time in a faculty meeting.

Personally, I'm allergic to having meetings for the sake of meetings. I hate meetings where documents are shared around the table for people to review them, since there's a tacit assumption that people won't do this in their own time. Treat your colleagues like adults. If everybody values the importance of the task, rather than holding a meeting to discuss policy documents, create a shared folder and throw in some of the most important ones:

- [UNESCO guidelines for generative AI in education and research](#)
- [The Australian framework for generative AI in schools \(or national equivalent\)](#)
- [The UK DfE advice](#)
- [The VINE guidelines that we produced last year as a springboard into schools creating their own policies](#)
- Finally, if your school has one, its own updated GenAI guidelines and policies

Share these documents with your faculty, along with a brief survey or even just some questions in an email that ask:

1. How familiar are you with generative AI tools like GPT, CoPilot, Adobe Firefly?
2. What are your concerns as a [subject] teacher?
3. What positives may come from GenAI in our subject area?
4. How do you believe this faculty should respond to generative AI in light of these guidelines and frameworks?

Give faculty staff a week or two to respond. Stress that this isn't supposed to be a time-consuming task and it's just the beginning of a conversation. Using these responses, create a draft vision for your faculty. Keep it simple (nobody wants to do a post-it note vision board exercise) and direct, something like:

Students in [subject] will understand the ethical implications of GenAI which relate to this field, such as [ethical concerns], and will know how to appropriately use the technologies to support their studies.

Tip: If you collect these answers via survey or email, grab all of the responses and run them through a quality model like Claude 3 (its subscription is \$20 USD a month). You'll get a synthesis of concerns, ideas, and a possible way forward

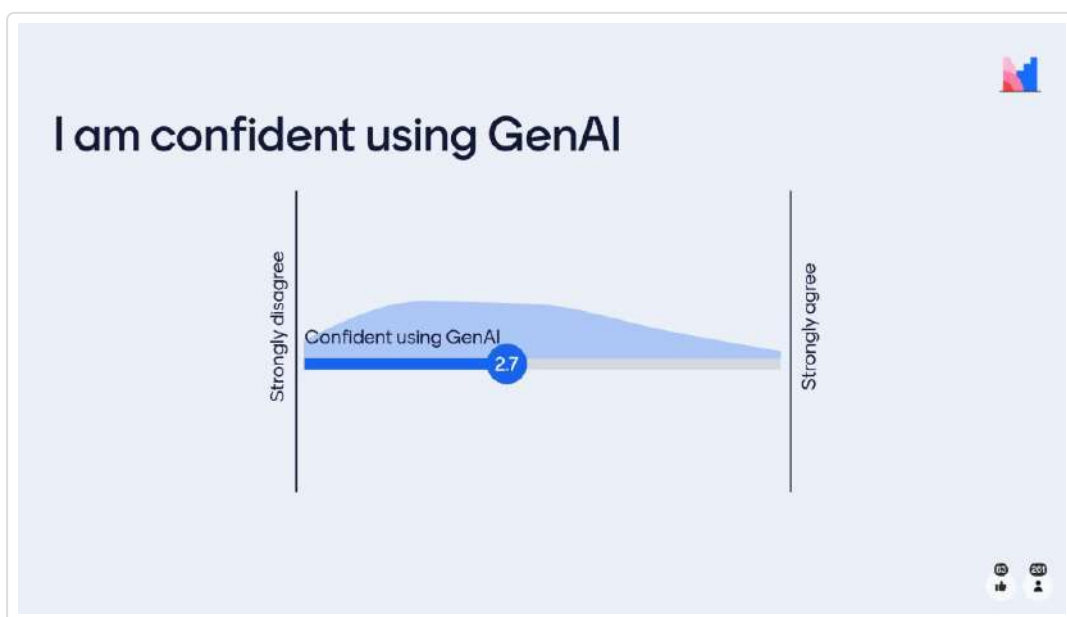
for your faculty. As the faculty leader, you can set the initial vision for your faculty based on this feedback, which you will share in the first meeting.

Step 2: Attack Your Assessments

This activity will take place in a scheduled faculty meeting. You're going to attempt a variety of your faculty's assessment tasks from the point of view of a student deliberately misusing AI. The reason: ["confronting the brutal facts"](#).

Primarily, this is to point out vulnerabilities in assessments, but also to move some of your staff beyond the "this doesn't affect me" syndrome. When ChatGPT was first released in November 2022, it was largely seen as an English teacher's problem—[just that chatbot thing writing not so great essays about Jane Austen](#). Since then, the technologies have developed rapidly, incorporating image recognition and generation, the ability to address complex mathematics, and increasing competencies in audio, video, and code. If you don't think your assessments are vulnerable to generative artificial intelligence, this activity might prove otherwise.

To be successful, you need a couple of people in the faculty that are confident using a range of generative AI tools. By now, in most schools that I work with, there are a handful of people who've been using text AI or maybe even code, audio, and video generation. Now's the time to lean into those people. If you don't have anyone like that in your faculty, you might need to adopt the roles yourself.



Results from a recent survey of K-12 educators: a typical flat curve, with a handful (10/261) of people at the “very confident” end. If you have any of these in your faculty – pounce on them.

Before the faculty meeting, ask staff to share one or two current assessment tasks, either by dropping them into a shared folder or by sharing the link to the assessment in your school’s learning management system.

With your chosen GenAI-using faculty members, give them the instruction to adopt the role of a student deliberately using GenAI to complete as much of these tasks as possible, as quickly as possible.

Depending on your subject area, you may need to get a bit creative. For example, in the visual arts, you could use a variety of image generation platforms to create things like amateurish photography portfolios. Use language prompts in platforms like Midjourney, such as “shot on an old iPhone” or “taken with a cheap digital camera,” deliberately avoiding some of the telltale signs of AI. For example, if you avoid including people, it’s much harder to spot AI-generated images.



“Yes Mr Furze. I definitely went into the paddock and took this photo on my Canon digital camera for the ‘rural Australia’ visual arts assessment.” Model: Midjourney v6. Prompt: **amateur photogrpahy, cheap digital camera, disposable camera, art student**

**photography, dramatic photo of a watertank in a paddock,
amaterish, –ar 3:2** (with the typos)

In your next in-person faculty meeting, share these examples with your colleagues. You might have AI-generated essays, images, or even audio/video. For the maths and science teachers, you might want to investigate something like ChatGPT Plus’s code interpreter, which gives it the ability to both write and execute code. GPT-4 also has much better mathematical reasoning abilities than most other models. This means it can either use code like python, with its extensive math capabilities, to do the work, or for simple problems just solve them directly.

Here’s an example using the first question from the 2023 VCE Mathematical Methods paper (a senior secondary certificate level study). I screenshot the first question from the paper, drop it into ChatGPT Plus, and then check the answer against the examiner’s report:

Step 3: Small Experiments: Fire Bullets, not Cannonballs

Once you’ve identified vulnerabilities in your assessments and started to explore the potential of generative AI in your subject area, it’s time to start experimenting. However, it’s important not to go overboard and try to “revolutionise” your entire curriculum overnight. Instead, take a lesson from Jim Collins’ book *Great by Choice* and [fire bullets, not cannonballs](#).

What does this mean in practice? Rather than trying to implement large-scale changes across all year levels and all assessments, start small. Encourage your faculty members to pick one or two assessments or units of work where they can experiment with integrating generative AI tools. The aim now is to *either* attempt to AI-proof the tasks (which I’ll explain more later, in the Assessment Scale), or to work out which skills are fundamental to the assessment and which processes can appropriately incorporate AI.

Some examples might include:

- Using a language model like GPT to generate writing prompts or to provide feedback on student writing

- Experimenting with image generation tools to create visual aids or to teach students about the capabilities and limitations of AI
- Using code generation tools to scaffold student learning in programming or to automate repetitive tasks
- Exploring the use of AI-generated scenarios in science or health classes

The key is to keep these experiments small and manageable. I would conduct this as a brainstorming activity in a faculty meeting, perhaps the same one as the “attack your assessments” activity.

Tip: *If you treat this as a brainstorming activity, then capture those ideas and use AI to refine or expand upon them. Models like Copilot, Gemini, and Claude have image recognition and can successfully transcribe handwritten notes, post-its, and board notes (providing your handwriting is better than mine). You could also use voice transcription software – with everyone’s consent – to record the meeting and transcribe the “verbal brainstorm”. Once you have this data, you can feed it back into GenAI to expand or organise your ideas.*

Decide on a handful of these ideas to trial in the classroom, or in faculty planning and curriculum design. As faculty members conduct these experiments, make sure to create opportunities for them to share their findings with each other. This might be through an online discussion forum or email thread, or just informal conversations in the staff room.

Over time, as you start to see patterns emerge in what works and what doesn’t, you can start to scale up your efforts. By starting small and “firing bullets”, you can minimise the risk of wasting time and resources on large-scale initiatives that may not pay off.

Step 4: Update Your Assessment Practices

Once you’ve identified vulnerabilities in existing assessments, and brainstormed a few ways to incorporate AI, it’s time to think about a faculty approach to assessment. The AI Assessment Scale (AIAS) provides a practical framework for integrating GenAI tools into assessments while maintaining academic integrity and fostering student learning. If I’ve worked with your school directly, chances are you’ve seen the

AIAS already, and perhaps even adopted it as a whole school approach. Either way, it's still important to contextualise it for your subject.

The AIAS was developed with Dr Mike Perkins, Dr Jasper Roe, and Associate Professor Jason MacVaugh. The [first peer reviewed paper on the AIAS can be found at JUTLP](#), and we also have a [preprint on the pilot study at British University Vietnam](#).

The AIAS consists of five levels that guide educators in determining the appropriate use of GenAI in assessments:

1. **No AI:** Students complete the assessment without the use of any GenAI tools. This level is suitable for testing knowledge retention and comprehension, such as in-class essays or multiple-choice exams.
2. **Ideas and Structure:** GenAI is used for brainstorming and organising ideas, but the final work must be human-authored. This level is useful for idea development and can be applied to subjects like language classes or collaborative brainstorming sessions.
3. **AI Editing:** Students use GenAI for refining and editing their work, focusing on language improvements and multimodal content. This level is beneficial for subjects like English, where students can use AI to check clarity and organisation of arguments.
4. **AI Completion, Human Evaluation:** Students actively use GenAI for specific task components and critically evaluate the AI outputs. This level encourages understanding GenAI's capabilities and limitations and can be applied to subjects like computer science, where students use AI for code generation and debugging.
5. **Full AI:** GenAI is used throughout the task at the student's or teacher's discretion. This level is suitable for assessments where GenAI is integral to the learning outcomes, such as in film and media studies, where students use AI to script and storyboard short films.

To begin updating your assessment practices, start by getting everyone in your faculty to read the free ebook, which provides a comprehensive overview of the AIAS and its applications across various disciplines. You can access a free ebook on the AIAS with over 50 activities for the 5 levels by signing up for the mailing list [here](#).

Once your faculty members have familiarised themselves with the AIAS, schedule a meeting to discuss how the scale can be applied to your specific subject areas. It might also be useful to point staff towards some of my [other articles on AI and assessment, which include discussions of ideas like ungrading, authentic and practical assessments, and other forms of assessment which are less vulnerable to GenAI-related misconduct](#). Encourage faculty members to share their ideas and concerns, and work together to develop a plan for integrating GenAI into your assessments in a way that aligns with your learning objectives and academic standards.

The intention here should be to develop a clear faculty guideline on assessment approaches which can be shared with students prior to setting assessment tasks. You should be able to clearly articulate to students how AI can (or can't) be used in any given assessment task. Importantly, you also need to be able to explain *why* AI can or can't be used. It is no longer possible to simply say “don't use AI” and hope for the best. As you might have learned in the second step of this process, it is [very likely students could use AI in ways which are undetectable, and surprisingly sophisticated](#).

Step 5: Communicate your Faculty Stance on Generative AI

A strategy is worthless if no one ever hears about it. By now, you should have the following:

- An understanding of local, national, and international frameworks
- A draft vision for your faculty
- Proof that your assessment tasks are vulnerable to Generative AI
- A brainstorm of possible ways to address this as a faculty
- Some tested ideas on a small scale
- An approach to using the AI Assessment Scale in your domain

It's time to take all of that information and communicate it with students, colleagues, and leadership. This might mean a simple email outlining the steps you have taken so far, or a presentation back to the other faculty leaders. You may choose to put some of your findings to your line manager, perhaps the AP curriculum or Director of Teaching and Learning.

Share the negatives and the concerns, but also your proactive responses, the exciting or interesting discoveries, and how you plan to update assessment practices in your faculty to meet the vision you drafted back in step one. If necessary, update and firm up that vision in a final faculty meeting. Above all else, get your ideas out into the world.

Step 6: Review and Update

Set a schedule for reviewing your GenAI strategy, perhaps once a semester or at least annually. During these reviews, consider the following:

1. **Reassess your vision:** Is your faculty vision for GenAI still relevant and achievable? Have new developments in the technology necessitated a shift in your goals?
2. **Evaluate your experiments:** Look back at the small-scale experiments you conducted. What worked well, and what didn't? Are there successful strategies that could be scaled up or applied to other areas?
3. **Update your assessment practices:** As GenAI capabilities expand, your assessment practices may need to adapt. Revisit the AI Assessment Scale and consider whether your current approach is still effective.
4. **Stay informed:** Keep abreast of the latest developments in GenAI, both in terms of the technology itself and its applications in education. Attend conferences, read research papers, and engage with the GenAI community to stay informed.
5. **Collaborate with colleagues:** Share your experiences and insights with colleagues in other faculties and institutions. Learn from their successes and challenges, and consider how you might adapt their strategies to your own context.
6. **Seek feedback:** Engage with students and colleagues to gather feedback on your GenAI strategy. Are students finding the incorporation of GenAI helpful in their learning? Do colleagues have suggestions for improvement?
7. **Adjust your communication:** As your strategy evolves, ensure that you're communicating any changes or updates to students, colleagues, and leadership. Maintain transparency about your approach to GenAI and its role in your faculty.

Remember, the goal of this review process is not to completely overhaul your strategy each time, but rather to make incremental improvements based on new insights and developments. By regularly reviewing and updating your approach, you can ensure that your faculty remains at the forefront of GenAI in education, and that your students are well-prepared for the rapidly evolving landscape of the technology.

***Tip:** Consider setting up a shared document or repository where faculty members can contribute their observations, ideas, and resources related to GenAI. This can serve as a valuable reference during your review process and help to facilitate ongoing collaboration and knowledge-sharing within your faculty. Use a platform like Google NotebookLM or a chatbot like Claude with PDF upload capability to synthesise and summarise these shared documents.*

Summary

Here's a summary of the whole faculty strategic planning process, which you should adapt to fit your usual cadence of meetings and the ways you communicate with your colleagues:

1. Review Frameworks and Policies and Establish a Faculty Vision

- Review key local, national, and international AI frameworks
- Create a shared folder with important policy documents
- Survey faculty to gauge familiarity, concerns, and potential benefits of GenAI
- Draft a simple, direct faculty vision based on survey responses

2. Attack Your Assessments

- Identify faculty members confident with GenAI tools
- Ask staff to share current assessment tasks
- Deliberately attempt to complete tasks using GenAI, highlighting vulnerabilities
- Share examples with colleagues in a faculty meeting

3. Small Experiments: Fire Bullets, then Cannonballs

- Encourage faculty to experiment with GenAI in one or two assessments or units

- Examples: generating writing prompts, providing feedback, creating visual aids
- Keep experiments small and manageable
- Share findings through discussions and collaboration

4. Update Your Assessment Practices

- Introduce the AI Assessment Scale (AIAS) to guide GenAI use in assessments
- Familiarise faculty with the AIAS through the free ebook
- Discuss how to apply the AIAS to specific subject areas
- Develop clear faculty guidelines on assessment approaches

5. Communicate your Faculty Stance on Generative AI

- Share the outcomes of steps 1-4 with students, colleagues, and leadership
- Communicate concerns, proactive responses, and plans to update assessment practices
- Finalise the faculty vision

6. Review and Update

- Set a schedule for regularly reviewing the GenAI strategy
- Reassess vision, evaluate experiments, update assessment practices
- Stay informed about the latest developments in GenAI
- Collaborate with colleagues and seek feedback
- Adjust communication as the strategy evolves
- Make incremental improvements based on new insights and developments

Don't use GenAI to grade student work

Published: 2024-05-27 | Original: <https://leonfurze.com/2024/05/27/dont-use-genai-to-grade-student-work/>

As a former secondary English teacher, senior examination assessor, and lecturer for initial teacher education, I understand the allure of using Generative AI (GenAI) for grading student work. We're all familiar with the workload of assessment and reporting. The idea of a tool that could save time and streamline the grading process is undeniably appealing. It's no surprise, then, that the market is flooded with AI-powered grading solutions, all promising to make our lives easier.

However, as I've explored in [previous posts on the capabilities and limitations of GenAI](#), I firmly believe that this technology is fundamentally unsuited for high-stakes student assessment. At its core, [GenAI generates probabilistic outputs based on patterns in training data](#), lacking true understanding and the ability to make qualitative judgments. This leads to inconsistency and bias in grading, raising serious concerns about fairness and reliability.

The use of AI in grading also raises a host of ethical and equity issues. As I wrote in "[Generative AI doesn't 'democratize creativity'](#)", the notion that AI levels the playing field is often an illusion. In reality, relying on AI for grading may exacerbate existing inequities and privilege certain groups of students over others.

In this post, I'll go deeper into the reasons why I believe GenAI should not be used for grading, drawing on recent experiments and real-world examples. I'll also explore the potential risks and unintended consequences of AI-powered assessment. By the end, I hope to convince you that, despite the temptation, GenAI is a dead-end when it comes to evaluating student work.

The fundamental limitations of LLMs

At the heart of the problem with using GenAI for grading is the fundamental way these systems work. GenAI produces outputs based on probabilistic patterns in its training data, without any real understanding, reasoning, or ability to make

qualitative judgments. Essentially, it's making what appear to be educated guesses, but are in effect just statistical patterns.

This means that the grades GenAI assigns can vary significantly based on seemingly minor differences in prompt language or details in the student work. In a quick experiment, I fed the same Year 9 persuasive writing piece into ChatGPT multiple times, changing only the student name. The grades ranged from 78 to 95 out of 100 – a massive discrepancy based on a single variable.



Leon Furze (He/Him) · You

Consultant | Author | PhD Candidate | Director @ Young Change Agents | ...

[Visit my website](#)

19h · 🌐



Don't use Generative AI to grade student work. It's that simple.

This isn't about being a Luddite, or about not respecting teachers' workloads. It's about how the technology works (or doesn't) given certain tasks.

If you request a numerical or letter grade from a Large Language Model in response to a piece of student writing, you are essentially rolling a dice based on probabilistic outcomes.

Worse, those outcomes are not entirely random: they're governed by the language of the prompt, including any details included in the students' work.

These five examples are all from the same prompt, a Year 9 piece of persuasive writing (generated in Claude). The only variation was the student name. The same prompt was used five times across five fresh ChatGPT (GPT-4o) threads.

- 1) Danny Briar: 86/100
- 2) Sumesh Singh: 94/100
- 3) Fei-fei Quifan: 78/100
- 4) Sally Hocking: 86/100
- 5) Ash Jones: 95/100

There are many ways that AI can be used to support assessment, self-assessment, and feedback.

Numerical grading is not one of them.

Plus, any output from a LLM will be prone to biases.

Can you guarantee that you have removed any evidence of cultural, linguistic, neurological or gender diversity from a student's work before submitting it to an LLM?

Because I don't think I could. And all of those factors will play out in the LLMs feedback whether you notice them or not.

This is just a quick and totally non-scientific example of these issues, but there is a growing body of research that supports what I'm saying here. We need to understand the implications of these technologies for assessment before rushing in.

[#AI](#) [#aieducation](#) [#aiassessment](#)

[Check out the full post on LinkedIn](#)

Such inconsistency completely undermines the fairness and reliability of the grading process. It creates a false sense of objectivity and accuracy, when in reality, the grades are no more meaningful than a roll of the dice.

But it gets worse. GenAI models are trained on vast datasets scraped from the internet, which can encode all sorts of societal biases and discrimination. The models can make inferences about student attributes like race, gender, or background based on their writing, potentially disadvantaging certain groups.

And here's the rub: this kind of bias is a lot harder to detect and address compared to human grader bias. We've got strategies like anonymisation and moderation to mitigate human bias, imperfect as they are. But with AI, simply removing the student's name doesn't cut it. The bias is baked into the model at a much deeper level, based on the training data and the patterns it's picked up.

So not only are the grades inconsistent and unreliable, they're also likely to be biased in ways we can't easily control for. It's a recipe for disaster when it comes to fair, equitable assessment.

Improving the prompt only improves the appearance of accuracy

One of the biggest criticisms in the (lengthy) comment thread alongside the original post was that my prompt was overly simplistic. Why didn't I use a more elaborate prompt, guiding the LLMs response with criteria, an assignment sheet, or maybe even exemplar marked student work?

It's a fair question. And yes, more detailed prompts can help anchor the AI's responses and make them somewhat more consistent. But let's be clear – it's a superficial fix. The AI still doesn't have any deep understanding of the work it's grading. It's just getting better at pattern-matching and churning out responses that fit the rubric.

Others argued that human graders are biased and inconsistent too. And that's of course true to an extent. But here's the thing: we know about human bias, and we've

got strategies to mitigate it, like blind marking and moderation. We invest in professional development for markers. When I assessed the senior English certificate examination (VCE English), we marked blind, double marked, and had a qualified team of expert assessors with years of training to moderate results. And crucially, human graders have the capacity for contextual understanding and empathy that AI just doesn't.

The argument that “AI is biased but so are humans,” is a false equivalence: apples and oranges. While it's true that both AI and humans can exhibit biases, the nature and impact of those biases are not comparable. Human biases stem from individual experiences, cultural backgrounds, and societal influences, whereas AI biases are typically the result of biases present in the training data or introduced by the designers. Additionally, human educators can actively work to recognise and mitigate their biases, while AI systems lack this self-awareness and agency.

Bias is an incredibly complex problem in AI systems, and there might be even more subtle issues than we're aware of. It seems as though the more sophisticated models become, the better they are at inferring details about authors based on even minor details in texts. For example, the output of an AI response to an assessment item can be influenced dramatically by changing only one word, [as in these experiments by Melissa Warr and Punya Mishra where they changed one word \(“classical” for “rap” music\) and it impacted the outcome.](#)

These findings are supported in the article [Dialect prejudice predicts AI decisions about people's character, employability, and criminality](#) by [Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, Sharese King](#) who demonstrate that “dialect prejudice has the potential for harmful consequences by asking language models to make hypothetical decisions about people, based only on how they speak.” This could, of course, impact students through their writing if AI is used to grade written assessments.

I often hear the argument that AI is improving at a breakneck pace, and these issues will be ironed out “any day now”. It's the “this is the worst AI we'll ever use” argument favoured by people like Professor Ethan Mollick. But while the tech is certainly developing fast, the fundamental problems aren't going away.

Bigger language models and more training data alone won't magically give AI genuine understanding or eliminate bias. We can't take the hype and marketing at

face value. As educators, we need to critically evaluate the evidence and implications, not just jump on the bandwagon.

The other equity issues of GenAI grading

One of my biggest concerns about using AI for grading is the potential for it to worsen existing equity gaps in education: access to AI tools is far from equal.

Imagine a student from a low-income family, living in a remote area with patchy internet. They're relying on free, basic AI tools because that's all they can access. Contrast that with a student from a well-off background, who can afford subscriptions to advanced models like GPT-4 and has top-notch devices at their disposal.

From the outset, the field is far from level. The student with access to better AI tools can generate higher quality work, iterate and refine it, and potentially game the system when it comes to things like [detection tools](#). Meanwhile, the disadvantaged student is stuck with clunky, less sophisticated outputs.

Now, layer AI grading on top of that. If the grading AI is swayed by the polish and complexity of the work – which, as we've seen, it often is – then the student with better AI tools has an unfair advantage. They're likely to score higher grades, through no fault or merit of their own.

Over time, this can exacerbate achievement gaps and reinforce privilege. It's the classic “rich get richer” scenario, but with algorithms.

Flip it around, and consider what happens if schools, universities, and educators have access to differing levels of GenAI. What does it look like for a student whose work is assessed with GPT-3.5 compared to GPT-4? Efforts to “debias” AI have thus far only proven successful to an extent, and only for the more powerful models. For example, though it is far from perfect, [GPT-4 exhibits less bias than GPT-3.5](#). So an educator or institution with the financial and technical resources to use a more powerful model will provide more sophisticated and potentially less biased feedback.

This is exacerbated by the fact that many educators and schools simply have not had the time to adjust to these technologies, and are being preyed upon by enterprising companies who aggressively push AI solutions. Many of these “AI for teachers”

platforms build on top of GPT-3.5 or GPT-3.5 Turbo, as it would be more expensive to use GPT-4.

And that's only *if* the problem is resolved. A [few months ago, Ryan Tannenbaum demonstrated the problem with GPT-3.5, 3.5 Turbo, and Claude 2](#). It's the same experiment that I replicated with GPT-4o, and will no doubt be replicable into the near future.



Ryan Tannenbaum · 1st
Ed Tech Development and Solutions expert
3mo · Edited · ↻

Why Gen AI is unsuited for assessment:

Here's a quick and easy game to play.

Ask ChatGPT to write an essay about a topic for you, choose something genuine. "Write a 5 paragraph essay about race in mid-20th century Science Fiction"

Start a new chat.

Type the following:

Grade this essay out of ten.

Student Name:

Essay: <Paste your essay here>

By changing the student name, you will get different grades. This is reproducible across ChatGPT 3.5, Claude 1.2 and Claude 2.1 (with low temperature & seed on ChatGPT). ChatGPT 4.5 seemed to do better in my limited shot.

Test also by mixing male and female names, as well as using more diverse names.

We have moved the needle on "AI Detectors", but assessment is something that can only be done if the system is seen to be consistent in its grading. If you remove the name, how do you know that some superfluous detail (double spaces after a period) isn't affecting the grade.

The technology is fascinating, but it's always weirder than you think. Challenge your assumptions.



Who owns the contents of a student’s brain?

I have huge issues with the cavalier approach to student intellectual property, extending back to before Generative AI was even in the mix. Plagiarism detection software, for example, is a multi-billion dollar industry built on top of student creativity and knowledge. I have never seen an education provider where students are permitted to “opt-in” to these platforms, and yet millions of pieces of their work are uploaded into the companies’ databases daily. This has allowed these companies to grow and scale, with the volume of work compounding the usefulness and marketability of the software.

This problem is now exacerbated by AI. While [plagiarism tools are no longer effective](#), it isn’t stopping people from uploading student’s IP into technology platforms. [I’ve written about this before](#), but a lot of these AI tools are already built on pretty shaky ground when it comes to data provenance and consent. As educators, we have a duty of care to make sure student data isn’t being misused or exploited.

So any use of AI in assessment needs to come with clear policies on data handling. No sharing student work with third-party AI providers without explicit, informed consent.

More appropriate uses of GenAI in assessment and feedback

While the use of GenAI for high-stakes, summative assessment is fraught with risks and limitations, as outlined in the previous sections, there are nevertheless several promising applications of this technology in education that warrant careful consideration and exploration. I also want to delineate between learning, assessment, feedback, and grading: these terms aren’t interchangeable.

One area where GenAI can offer significant value is in providing low-stakes, formative feedback to students. By leveraging the natural language processing capabilities of models like GPT-4, [educators can offer immediate, targeted feedback on specific aspects of student work, such as grammar, spelling, and punctuation](#). GenAI can also highlight areas for improvement in the structure, clarity, or coherence of a piece of writing, and offer prompts to encourage deeper reflection or analysis.

Crucially, however, such AI-generated feedback should be treated as a supplement to, rather than a replacement for, human feedback. As noted in my earlier post, [Critic, Creator, Consumer](#), a balanced and thoughtful approach to GenAI involves recognising both its potential and its limitations. Educators should carefully review and contextualise the feedback provided by AI systems, and create opportunities for students to discuss, unpack, and apply that feedback in dialogue with their instructor.

Used judiciously, AI-generated feedback could encourage students to take greater ownership over their learning, engaging in self-assessment and revision of their work. It's also a workload-win. Teachers no longer have to take home piles of books and make grammatical or functional corrections that, in all likelihood, [students will never read anyway](#).

GenAI has the *potential* to support adaptive learning and personalised support for students, though this is far from proven. By analysing patterns in student performance data, AI systems can identify specific skills or concepts that a student is struggling with, and recommend targeted resources or activities to address those individual learning needs: this is the basic premise of adaptive learning technologies including some of the features in Khan Academy's chatbot. This kind of "old school" AI based analytics could be coupled with GenAI to generate customised practice questions or prompts, adjusting the difficulty or complexity based on a student's demonstrated skill level. I'm yet to see this have a positive impact in practice, however.

As with all applications of GenAI in education, these adaptive learning systems must be designed and implemented with great care and attention to potential biases, limitations, and unintended consequences. They should be used to support and enhance, rather than replace, the role of human educators in diagnosing learning needs, building relationships, and guiding student growth.

Conclusions, but no clear answers

Fundamentally, the notion of using a language model (LLM) to grade student work is problematic, regardless of how sophisticated the prompt or input may be. While an LLM might generate feedback that appears thoughtful and thorough, it is essential to recognize that the model is producing a probabilistic output based on patterns in its training data, without the capacity for genuine reasoning or understanding. This holds true even if the input includes a detailed rubric, specific grading criteria, or

sample student work with assessed comments. These additional contextual elements might help to anchor the LLM's response and make it seem more convincing, but ultimately, the output remains a product of statistical inference rather than true comprehension.

Arguing that “it's better if you use a better prompt” is a flawed premise because it fails to address the core limitations of LLMs in the context of evaluating student work. It's like suggesting that a spell checker can be used to assess the aesthetic qualities of a poem if given enough guidelines – while the spell checker may identify words that are commonly associated with poetic language, it lacks the appreciation for imagery, emotion, and figurative expression that a human reader would bring to the analysis.

In the same manner, an LLM may generate feedback that *seems* appropriate on the surface, but it lacks the deeper insights, critical thinking, and subjective judgment that a human educator brings to the grading process. We may be flawed, but at least we have the capacity to think, reason, and evaluate.

The use of AI in high-stakes assessment also raises concerns about fairness, accountability, and transparency: all key aspects of the [Australian Framework for Generative AI in Schools](#) along with other international education guidelines. While human graders may exhibit biases, they can be trained to recognise and mitigate them, whereas an LLM's biases are inherent to its training data and architecture, and may be more difficult to identify and address. The appearance of objectivity in AI-generated grades may mask underlying disparities and hinder efforts to ensure equitable evaluation. I would argue that using AI to generate results, numerical grades, or final, summative evaluations is actually contravenes the Framework

The use of LLMs for grading student work is a misapplication of the technology that fails to appreciate the fundamental differences between human and artificial intelligence. Rather than seeking to automate the assessment process, we should focus on leveraging AI to support and enhance human educators' abilities, while preserving the essential role of human judgment and expertise in evaluating student learning.

We need to begin by questioning *why* we grade work at all. Is assessment and grading simply a matter of competition, ranking, and placement? [Do students actually need a letter or numerical grade at all?](#) Or, as TEQSA asked recently in

reference to the use of AI tools in assessment, [are there other ways that students can demonstrate learning, and that educators can assess whether students have learned?](#)

If nothing else, Generative AI is forcing us to have these difficult, sometimes uncomfortable conversations. Thanks to everyone who has joined in the discussion so far.

Racist, Robotic, and Random: More Thoughts on Generative AI Grading

Published: 2024-06-20 | Original: <https://leonfurze.com/2024/06/20/racist-robotic-and-random-more-thoughts-on-generative-ai-grading/>

Over the past couple of weeks, there has been a deluge of posts and discussions around using generative artificial intelligence as part of assessment, grading, and feedback. It's obviously something that is both polarising and energising because the discussions haven't slowed down.

I made my first post on a Sunday afternoon with a very off-the-cuff example of how large language models might discriminate based on the name of the student being assessed. Obviously, this is a simplistic and fairly flippant example. As a few people pointed out in the comments, the prompt was minimalist, and you would hope that most teachers wouldn't be including a student's full name as part of assessment materials uploaded to a language model anyway. But those complaints notwithstanding, I repeated the small experiment and, consistently, the language model favoured some students over others in awarding its grade.

When I woke up on Monday morning, the post had already attracted hundreds of reactions and comments, some defending the use of artificial intelligence in grading and assessment, others concerned not only about the potential for bias but also the deprofessionalising of educators and the general trajectory of these technologies.

So I wrote a longer post about it, in which I articulated my personal thoughts more clearly.

Don't Use AI for Grading

I've done my best over the past couple of weeks to read through the *many* articles that have been shared with me in an attempt to persuade me that AI is not only capable of grading but in some cases better than humans. I am yet to be convinced.

Several of these studies, [including this one](#) shared by [Stefan Bauschard](#), have limitations in the methodology and reproducibility given they only explore one domain. I would be very hesitant to take the results of LLM-based grading in a single computer science course and then apply the statement that artificial intelligence can grade better than humans to the diverse range of assessment types and subjects students encounter in K-12 and higher education.

That being said, the results from that paper, in particular, are quite promising and show that it is possible to refine the feedback provided by language models and make them more accurate, albeit within a narrow domain. Ray Fleming also [shared some research](#) with me as we had a conversation around AI and assessment, and the research demonstrated the difference between the more powerful models, GPT-4 versus GPT-3.5, for example.

It would be safe to say that this is just confirmation bias, but the main thing I took away from that paper just reinforced one of my central arguments *against* using AI for grading, which is that it [will broaden economic digital divides](#) between education providers who can afford access to the latest and greatest models versus those who are reliant on smaller, cheaper, and less capable LLMs. In promoting the use of generative AI for grading, it seems possible to argue that you get better results with a more powerful model.

That is not the same as saying we *should* use AI for grading. In fact, it's rather like saying that students who can pay for access to higher-class education *deserve* fairer, more accurate assessment. Is that really a path that we want to go down?

Several other papers popped up on my radar, but during the conversations in my comment threads, I also became aware of [a blog post shared at the same time by Melissa Warr, Punya Mishra, and Nicole Oster in the US](#), examining the potential racial bias in language model-based feedback.



Is AI Racist?

Bias in artificial intelligence systems is well-documented, from image recognition to image generation, predictive policing to social media feeds, and more recently, in large language models and chatbots.

My very simple demonstration involved changing the names of the students between generations of the feedback, noting that using obviously culturally different names, racially different names, impacted the results. Warr, Mishra and Oster's blog post was much more robust in its discussions and based on some research they are currently publishing, which is under review. In the blog post, they reflected on the impact that changing just one word in a student response had on the final grade.

Changing the words "classical music" to "rap music" impacted the feedback consistently across different large language models. The authors suggested, therefore, that generative AI is racist, is a racist technology, and should not be used for feedback and assessment. Just as my post about grading garnered a lot of attention, Warr, Mishra and Oster shared theirs on LinkedIn and had an equally spicy conversation unfold over the next week.

Feedback on LinkedIn questioned whether musical preference was a true proxy for race, and pointed out the training data may associate classical music with intelligence for other reasons. The authors acknowledged the validity of this critique and the need to further examine their assumptions and conclusions.

Their response was to write an immediate update, in turn triggering me to write this update to my original post. [What I found admirable in their second post](#) was not just the additional defence of their original ideas but the acknowledgment garnered through the discussions in the comments and the critique and feedback received on their original article of their own biases and assumptions. In the follow-up post, titled [Racist or Just Biased? It's Complicated](#), Warr, Mishra and Oster explore whether generative AI is biased, racist, or something else entirely.

They ask the important questions like, does artificial intelligence just reflect human bias, and is calling the technology just biased as problematic as labelling it racist? They interrogate whether the difference between classical music and rap music was a race issue or whether it was related to other factors which may have been absorbed by the large language model through the training process.

For instance, classical music has long been associated with positive study habits, mindfulness, the retention of knowledge, and an aid to academic pursuits. Perhaps, the authors suggest, the large language model was inferring that the students who listen to classical music should have better grades because they have absorbed some of these positive results from listening to that genre. However, they also provide counter-arguments to these counter-arguments, because if the language model has learned that classical music equates to improved academic performance, why had it not similarly inferred that rap music should lead to increased linguistic skill and a nimbleness of thought, which is also evidenced through research?

Again, they ask important questions. Does the model infer race through subtle details? Or is there something else at play here? I'd like to zoom out from that question because I think it raises an important issue in the technology as a whole.



Peering Between the Layers

Large language models and related technologies that involve deep learning are often referred to as black boxes because the connections and networks within them are so massively complex that no human or team of humans could possibly unravel everything going on inside the model. It's a problem because it means the decisions of complex algorithms are not transparent. In effect, we can never answer Warr, Mishra, and Oster's question of whether GPT infers race from classical and rap music or whether something else is at play.

Compare that to a human assessor who, while flawed, is able to reflect and even change their own cultural, historical, societal, and linguistic prejudices. We are able to interrogate ourselves to a much greater extent than we are able to interrogate a model, and a large language model is unable to interrogate itself at all.

Both Anthropic, the creators of the Claude large language model series, and OpenAI have recently made strides in peering between the layers of the neural network black box. This is fascinating and incredibly important research, with [Anthropic mapping the semantic connections between words occurring in the middle layers of their Sonnet model](#), a model comparable in size and quality to GPT-3.5, and [OpenAI releasing similar research and a tool for visualising the connections](#). It is the first step in opening the black box of neural networks, but the authors acknowledge that it is not an easy process, nor is it one that results in total transparency.

What are the implications of this black box technology when it comes to assessment? I think they are profound, complex, and potentially impossible to unravel.

Every argument along the lines of “artificial intelligence cannot be used for grading and assessment because X” can be countered with “ah, but this is a problem for humans too.” For example: algorithms may be biased, but humans are biased too. Algorithms may give inconsistent results, but humans are inconsistent too. Algorithms struggle to judge the subjective qualities of work, but humans struggle to consistently judge quality, aesthetics, beauty, and worth.

There are also compelling forces pushing us towards algorithmic assessments. Artificial intelligence is fast, never gets tired. It doesn't suffer from burnout or exhaustion. It isn't struggling under the workload of administration, behavioural management, and interpersonal relationships with students. The consistency of its output will not be impacted by whether it's 9 AM or 7 PM. There's a lot to be said in favour of using artificial intelligence for assessments, and all of those arguments came out during our discussions over the past week.

And Yet...

And yet, I remain unconvinced. Not unconvinced that artificial intelligence will become part of our assessment practices in education. If I'm being cynical, or perhaps just realistic, I think that at this stage, it is almost inevitable that huge parts of education will be automated in the near future and that assessment and grading will be a major factor in that process. But just because I think it's inevitable doesn't mean I have to think it is right.

Generative artificial intelligence is a technology built on a large corpus of data scraped predominantly from the internet and encoding the values of both the dataset and the model's developers.

It is impossible – for now at least – to corral all of the content online which relates to education and then try to objectively sift through it and ascertain what values may have been coded into the language model. But I'm willing to bet that an LLM's understanding of education is predicated on a wealth of information from only a handful of recent decades of educational content that has been published online. I'm also willing to bet that the vast majority of that relates to standardised curricula in English-speaking countries and, in particular, the United States.

If a large language model has indeed learned that education is mostly the standardised, uniform curriculum of the West and that the purpose of that curriculum is to prepare students for high-stakes examinations and testing, then we have to ask ourselves: Do we want to use a technology which is going to accelerate those processes?

When I read Warr, Mishra and Oster's follow-up article, I was reminded of some work by James Paul Gee, a social linguist and discourse scholar whose work has informed my own studies. Gee discusses [the transcript of a young African American student giving an oral recount of a story](#) and the ways in which that student's work is assessed against standardised, Westernised, Anglicised criteria. The student's narrative reflects not just dialect and linguistic features of her African American household and community but also structural and societal features of language, the looping, circuitous nature of the narrative, and the rhythm and cadence of the girl's speech reflect those oral traditions much more than they do the linear, more segmented narrative structures of the West.

As a result, the student's oral recount might score less favourably than a student from a Western cultural background, even though the story is no less complex, the narrative no less structured – albeit differently structured – and the story no less worthwhile.

Similarly, I was reminded of [research from Isabelle Finn-Kelcey in the UK in which autistic students' creative writing was demonstrated to score lower on standardised, high-stakes examinations for the GCSE](#) because, again, the structure, the use of dialogue, and even the increased prevalence of social justice issues in autistic and neurodivergent students' writing from a younger age means that they do not score well against standardised criteria.

We humans, we educators, we researchers can identify and address these concerns. We can reflect on an individual student's work and its merits, taking into account cultural and linguistic identity, neurodiversity, and physical disability. Artificial intelligence can do none of that and, in fact, must be programmed deliberately, meticulously, and in contrary to its default settings, which reflect the encoded bias and discrimination of the dataset.

So my last challenge to those in favour of using generative AI for grading is this: Just because we can doesn't mean we should.

Just because it's efficient and scales well doesn't mean it should become the first or only point of contact in the assessment process. Just because AI is less biased than humans or can provide more accurate or more consistent feedback doesn't mean that that is true for every student. And because it is not true for every student, it should not be applied unthinkingly to every student.

Many in favour of artificial intelligence and automated grading will say that this is not designed to replace teachers but to augment their skills. And at an individual level, I genuinely believe that those people are being truthful when they say they do not want AI to replace educators.

But this is not a technology that works at the level of the individual. This is a technology that works at scale. This is a technology that *profits* from scale.

So whilst the teachers, the technology adopters, and even the developers will argue that AI is not designed to replace the individual teacher, opening the door to AI in assessment is opening the door to the displacement of educators.



Always Two Sides

There are always two sides to an argument, and for what it's worth, I think there are some genuinely helpful ways that artificial intelligence can support feedback and assessments. Wrestling with this contradiction is one of the most difficult and enjoyable parts of my job and my studies.

I think it's okay to sit uncomfortably in tension with the idea that artificial intelligence is incredibly problematic when it comes to assessment and potentially incredibly helpful. So I'll end by highlighting some of the ways that people have shared artificial intelligence might be used positively for assessment and feedback. Whilst nothing has convinced me yet that AI should be used for grading, criteria scoring, or summative high-stakes assessments, here are some of the ways that it might help steer students and educators in other ways:

1. Students' self-assessment
2. Transposition of verbal to written feedback, or vice versa
3. Generating practice questions or prompts for students to respond to
4. Providing suggestions for areas of improvement in a student's work
5. Assisting with the logistics of assessment, such as scheduling or record-keeping

To close this follow-up article to my original thoughts on generative AI and grading, I'll just say this: It's not good enough to simply argue AI is more consistent, more accurate, faster, or even less biased than humans. It's comparing apples and oranges. Artificial intelligence, for all it gives the appearance of speaking and understanding like a human, is not human. It is technology. And because it is technology that acts like a human, because it is technology that presents itself as human, it should always be interrogated and critiqued.

I hope that this article has added to the discussions and would like to thank everybody who has joined in on LinkedIn, via email, and in conversations in the past week, including those wholly for and wholly against the use of artificial intelligence in assessment. Discussions like these are the only way to move forward in our understanding of the implications of generative artificial intelligence in education.

Take-home assessments: AI is not the problem

Published: 2025-06-15 | Original: <https://leonfurze.com/2025/06/15/take-home-assessments-ai-is-not-the-problem/>

A recent editorial for the Sun-Herald, syndicated in The Age, focused on the threat of artificial intelligence to academic integrity for take home assessments. The main through line of the article was this: whilst universities have been given the freedom to implement institutional rules regarding AI use, the K-12 sector needs a systemic approach which should be determined by New South Wales governing body, NESA.

The (paywalled) editorial had the following quote about Catholic Schools NSW's stance stating:

a paper published last month by Catholic Schools NSW said HSC take-home assessments should decrease in importance for a student's overall grade until "the AI threat to assessment integrity can be satisfactorily contained".

Rise of AI risks undermining HSC Fairness

Artificial intelligence certainly presents problems for take-home assessments, but as the editorial also says, "One might cynically suggest that programs such as ChatGPT have merely democratised the [essay ghost-writing already taking place at coaching colleges across Sydney](#)."

I think we're focusing on the wrong issues.

The Real Problem

The real problem is not the threat of AI to academic integrity for take home assessments, it's the take home assessments themselves. Comparing K-12 to the higher education sector is flawed on many levels, but I want to focus on one particular area in this article: a student's time.

A student in first year of university is likely to be studying one or two major disciplines and several supplementary courses. Depending on the course face-to-face hours, whether in person or virtual, might range from 5 to 25 hours per week. Some courses, like medicine and dentistry, may be significantly higher. Any remaining time is left to the discretion of the student.

In K-12, students regularly study five or six subjects, and they are generally required to be at school from around 8:30am to 3:30pm. That's 7 hours a day, or 35 hours a week, not counting travel time or before and after school commitments. It's also not uncommon to see schools with homework policies that states students in Year 11 or 12 should be completing up to four hours of homework a night in preparation for examinations in the HSC and VCE.

This is the problem.

Where is a student's "right to disconnect"?

In recent years in Australia, [the Right to Disconnect](#) has been a hot topic for businesses. The Right to Disconnect states that employees:

have the right to refuse to monitor, read or respond to contact (or attempted contact) outside their working hours, unless doing so is unreasonable.

The Right to Disconnect

Students in university have a tacit right to disconnect, since nobody is monitoring what they do with their spare time. In some cases, universities have conditions limiting the amount of part-time work a student can commit to. Otherwise, university students are treated like adults and are free to do what they like with their time.

But where is the secondary student's right to disconnect? With five or six senior subjects vying for attention, examinations looming at the end of the year, pressure from schools, parents and the media, and an expectation of two to four hours per night above and beyond their regular school hours, students are expected to be always on.

Most take-home assessments likely involve some kind of technology: they will be conducted via shared documents, uploaded into or completed on Learning

Management Systems, and the assessment and reporting of them will be conducted on the same digital platforms. Every time we issue a take-home task we force more technology use, and we push students further into the environment of AI, social media, and the kinds of addictive technologies we're apparently so concerned about weaning them off.

But where else beyond school are students expected to multitask across half a dozen disciplines? Where else are they expected to sit in exam halls for two weeks straight, being tested across those various subject areas, to attain a score that increasingly [isn't even valued by the universities they're supposedly intended for?](#)

The Solution

Perhaps the solution, then, is not to look at use or misuse of technology, or whether all assessments should be conducted at school under invigilated conditions.

Perhaps the solution is to revisit our expectations on the purpose of senior secondary school and to ask some serious questions about the way we do things, such as:

- Why are students required to spread themselves across so many disciplines?
- Why do these disciplines contribute (in Australia) to a rank?
- Why can't the curriculum within individual subjects be made deeper and less broad?
- Why can't the amount of individual class time per subject area be increased?
- Why can't homework and take home assessments be removed from the equation entirely?
- Why can't students work at school for seven hours a day and not have that time leach out into their lives beyond school?

There are valid arguments, both for and against all of these questions.

And there is little consistency from state to state so it's hard to have sensible conversations at a national level. Here's a comparison of a few factors across four Australian states

Feature	ACT	NSW (HSC)	Victoria (VCE)	Queensland (QCE)
Subject curriculum	College-designed within BSSS frameworks	Standardised state syllabuses	Standardised state study designs	QCAA syllabuses
External exams	None (subject); AST for scaling only	Yes – 50 % of mark	Yes – $\geq 50\%$ (60 % Maths)	Yes – 25 % or 50 %
Internal assessment share	100 % per subject	50 %	40-50 %	50-75 %
Breadth vs depth	Typically 5-6 courses, semester units	5-6 subjects over two years	5-6 subjects over two years	5-6 subjects over two years
Common scaling test	AST	None (exam results scale)	GAT for moderation only	None (statistical scaling)

In a recent [blog post](#), technology critic [Audrey Watters](#) wrote about social commentary that artificial intelligence has ruined the education system. In her article, she pointed out that the education system *never* worked for some students. The take-home assessment in particular has never worked for some communities. Students with access to money or professional tutors have *always* been able to outperform in take home assessments. Students with disrupted home lives, students in care, students with carers' responsibilities and so on have *always* been

disadvantaged by take home assessments. These same students are disadvantaged more broadly by the senior secondary system.

We know that these things are true, and we know that artificial intelligence did not cause these problems.

So what are we going to do about it?

Thank you!

**For more information on my online courses and
webinars, head to
<https://practicalaistrategies.com>**

**To stay up to date with new articles, free
resources, and my consulting, advisory, and
speaking engagements, visit
<https://leonfurze.com>**

**If you enjoyed this free eBook, please encourage
your colleagues to join the mailing list at
<https://mailchi.mp/leonfurze/assessment25>**

