Generative AI technologies and their role within assessment design

Michael Grove (Deputy Pro-Vice-Chancellor, Education Policy and Academic Standards, and Professor of Mathematics and Mathematics Education)
Contact: m.j.grove@bham.ac.uk

Introduction

Assessment serves multiple purposes within education (Brown, Bull and Pendlebury, 2013). One of its primary functions is to provide evidence of student learning that accurately reflects the efforts of an individual, and clearly demonstrates mastery of their knowledge, understanding, and skills. However, recent advances in, and the availability of, generative AI technologies mean that such tools now have the potential to be increasingly used by students for the completion of their assessments. This necessitates careful consideration of assessment design to ensure that where generative AI tools might be used by learners, their use is responsible, and the contribution of the individual learner to the work can be clearly demonstrated.

Importantly, consideration of the implications of generative AI technologies should not be used to automatically favour one assessment type over another, such as for example the increased use of invigilated on-campus examinations, which themselves have both benefits and limitations in assessing students' learning (Buckley, 2023). Instead, this offers a timely opportunity for academic members of staff to consider good assessment design and more broadly seek to answer the questions: Why are students are being assessed? To what extent are they being assessed? What is being assessed and why? This should be undertaken alongside carefully examining how evidence of individual student learning and the associated achievement of learning outcomes is being collected at a programme, rather than module, level.

Considerations for assessment design

Whilst academic staff may not be required to use generative AI technologies within their teaching, all must now consider the potential impact upon student learning and assessment.

This should be reviewed regularly given the pace at which these tools are developing and their likely increasing availability within other mainstream technologies. Additionally, higher education institutions are now establishing frameworks that require, or support, their students being provided with opportunities to engage with generative AI tools at a range of levels throughout their programme of study (University of Birmingham, 2024). This therefore presents both opportunities and challenges for the assessment process.

When considering the potential role of generative AI within assessment, a useful starting point is to determine whether:

Its use within an assessment is appropriate or encouraged. If so, tasks should be designed that promote or require the responsible use of generative AI tools by learners within the assessment process.

Its influence should be limited. That is, an assignment should be designed so that generative AI use by learners is less significant. Changes should focus upon either the nature of the assessment or the assessment process itself, and additionally by educating students on the potential implications of using generative AI tools upon the development of their own knowledge and skills.

It is not allowable. If the assessment requires the student to demonstrate only individual knowledge, understanding and skills, additional security measures are likely to be required to ensure that generative AI tools cannot be used.

Using generative AI technologies to review assessments

If an assessment has already been designed, it should be reviewed to explore what the potential implications of generative AI technologies might be. One approach is to use generative AI tools themselves.

A range of tools are available for this purpose. However, at some universities tools can be used where the transfer of user information and inputs beyond the institution is restricted. For example, at the University of Birmingham all staff and registered students have institutional access to Microsoft Copilot within Edge, a generative AI powered web chat tool that enables free access to GPT-4 and DALL-E 3 within a data protected environment

(University of Birmingham, 2024). Other generative AI tools can be used, such as Open AI's ChatGPT, but it is important users are aware that any data uploaded may not be secure and could be used to enhance or develop future iterations of these technologies (OpenAI, 2024).

The following approach can be used to explore the extent to which generative AI is able to answer, or provide insight into assessments:

- Paste the assignment text into the generative AI tool, checking the syntax, and see what it produces as an output.
- After reviewing the output, provide further text inputs to help fine tune the response. This 'prompt engineering' helps guide the model towards the kind of response being sought – the better the prompt, the more detailed and relevant the AI response will be.
- Try adding a range of further information: background or context to the assignment, sample work, marking guidance, or extracts from course notes, to see if this changes the output. Some AI tools allow you to upload whole documents directly.

It is important to review the final output carefully. Generative AI tools can produce responses that on first inspection appear to have the expected structure, but often lack essential detail or contain significant factual errors. Responses may also not be contextualised, for example to align with the required disciplinary focus of an assessment, or the output not prioritised upon the aspects that are most important, such as a critical analysis of key ideas as opposed to simply a presentation of the ideas.

It is also worth investing time in 'prompt engineering' (Liu, 2023). By understanding the possible generative AI output, the wording or emphasis of the assessment can be modified to focus upon aspects where generative AI tools are less successful in their response. Similarly, grading criteria should be updated for all assessments to reward the human elements required to ensure the submission is at the expected academic level.

The vulnerability of assessment to generative AI

Some assessment formats are inherently more vulnerable to the effects of generative AI than others. Table 1 shows a list of some common assessment types and highlights their potential vulnerabilities. It includes assessment types that are currently intrinsically more resilient to the effects of generative AI tools, and highlights where its current limitations might start to become apparent if used within each of the assessment types.

Assessment Type	Risk Level	Description and Identifiers
Non-invigilated quizzes and tests	Very High	Quizzes and tests are vulnerable where questions ask students to define or reproduce basic disciplinary knowledge. Generative AI tools can respond very effectively to multiple choice (closed response) questions based around factual recall or basic knowledge application and can also provide supporting options to explain why each option is correct or incorrect.
Complete at home short questions	Very High	Online non-invigilated examinations, which typically contain a significant proportion of short answer (open response) questions involving the recall of knowledge or basic knowledge application are also vulnerable. The ability of generative AI tools to provide responses in real-time also negates the effect of reducing the period within which the assessment is completed.
Essays	High	Non-invigilated essays on broad, general, and well-known concepts are especially vulnerable to the impacts of generative AI where it can excel at presenting information and mimicking writing styles. Whilst generative AI can develop essays that may appear consistent and follow a logical structure, they can fall short in key areas like developing strong independent arguments, analysing and evaluating evidence, establishing connections between ideas, and demonstrating original thought.
Technical reports	High	The risks associated with reports are similar to essays, particularly if focused upon well-known topics, examples or issues. The ability of generative AI tools to process large amounts of data can lead to reports that appear factually accurate but that often lack depth. They might plagiarise existing work by combining information from various sources without truly understanding the underlying

		concepts or properly citing them. An AI-generated report might present data but lack a clear explanation of its significance, the reasoning behind the methodologies used, or the limitations of the study and may struggle with the crucial aspects of analysis and interpretation. Some generative AI tools can, however, facilitate the direct analysis of a dataset or provide computing code.
Projects and dissertations	High	The potential risks of generative AI tools for projects and dissertations are similar to those for essays and technical reports. Risks can be mitigated by requiring a novel component, either disciplinary or localised, to the work. Most at risk are projects and dissertations that form literature reviews or summaries of well-known topic areas.
Reflective accounts	Medium	Reflective accounts or commentaries, whilst focusing on a student's personal learning journey, can be susceptible to manipulation by generative AI. One area of vulnerability lies in the potential for generative AI to mimic surface-level reflection. They can be trained on student reflection examples and may be able to generate text that uses appropriate vocabulary and references specific course content. However, genuine reflection requires introspection, self-evaluation, and a critical analysis of the learning process, aspects that generative AI currently struggles to replicate. Generative AI tools can fabricate experiences or learning outcomes, but whilst they can process course materials, they cannot replicate the actual experience of grappling with concepts, participating in discussions, or overcoming challenges. An inauthentic commentary might therefore present a somewhat sanitised version of a student's learning journey lacking the genuine struggles and growth that a student would typically be expected to describe.
In-person invigilated examinations	Low	In-person invigilated examinations, where access to third-party materials and online materials is typically restricted, are naturally more resilient to the potential impacts of generative AI tools than online non-invigilated assessments. However, where they include questions requiring the recall of knowledge or basic knowledge application, students might pre-generate and memorise responses to commonly used question types. The questions used

		within examinations should therefore be reviewed and refreshed on an annual basis.
Academic portfolios	Low	Unlike a single exam or essay, academic portfolios showcase a collection of student work that has been developed over time. This cumulative aspect makes it difficult for AI to develop a portfolio that is cohesive and reflects an individual student's learning journey and development. Further, portfolios often contain diverse materials like creative projects, drafts with revisions, and reflections. This variety challenges current generative AI tools which can struggle to adapt to different formats and content types in a coherent manner. Portfolios also often emphasise critical thinking skills like selection, curation, and self-reflection. These skills are not easily replicated by generative AI, which can struggle to explain the rationale behind the chosen materials or articulate genuine personal growth.
Scaffolded assessments	Low	Scaffolded assessment breaks down a complex learning objective into smaller, more manageable steps. A scaffolded assessment provides students with a staged series of tasks geared towards achieving an overall outcome. Each task is accompanied by instructions, support and measures to help check progress and enable the development of knowledge, understanding and skills. Unlike a single test where generative AI tools might mimic successfully the final answer, scaffolded assessments track progress over time. Drafts, revisions, reflections, and feedback from staff are all part of the evaluative process. These elements are challenging for AI to produce as they require genuine understanding and adaptation throughout the learning process. Further, scaffolded assessments emphasise critical thinking and problem-solving alongside the acquisition of knowledge and skills. It is difficult for generative AI to replicate the thought process behind a solution or the ability to learn from mistakes, skills that become evident through scaffolded tasks, and the interactions with staff members and fellow students.
Hybrid assessments	Low	The strength of hybrid assessments in their defence against generative AI tools lies in their diversity. By combining different formats and question styles, they present a more complex challenge for AI, which may struggle to adapt to open

		ended questions, essays, or practical tasks. Effective hybrid assessments move beyond rote memorisation, to emphasise higher-order thinking skills like analysis, problem-solving, and application of knowledge in new situations. The human component also has a critical role: hybrid assessments often incorporate elements that require human interaction and judgment, such as presentations, individual or panel discussions, or open-ended questioning.
Interviews or oral assessments	Low	Unlike written work, interviews and oral presentations rely on dynamic interaction and human judgment. This makes them more challenging for generative AI to exploit. Their key strength is the ability to enable follow-up questions and to engage the student in meaningful dialogue about their discipline area. Generative AI struggles to adapt to these dynamic exchanges and demonstrate genuine understanding. Furthermore, oral assessments can directly evaluate communication skills and critical thinking in real-time. These are areas where AI tools remain under development, making it difficult to convincingly replicate natural human communication or thought processes.
Practical assessments	Low	Practical assessments require students to apply their skills and knowledge in real-world settings. Generative AI struggles with tasks that demand physical manipulation, creativity, and real-time adaptation. Building a prototype, conducting an experiment, or performing a complex procedure all fall into this category. These hands-on activities require problem-solving, critical thinking, and on-the-go adjustments. Furthermore, practical assessments often encourage originality and showcase a student's unique approach. Generative AI, whose responses are based upon the datasets upon which it has been trained, finds it difficult to mimic this level of individual creativity and initiative.
Synoptic assessments	Low	Synoptic assessments, unlike traditional tests focused on a single topic, require students to demonstrate their ability to combine, understand and apply their knowledge and skills from across a discipline or a range of modules within their programme. Generative AI typically excels at specific tasks within a single domain, and the emphasis on connecting ideas across different areas and from perhaps diverse sources makes synoptic assessments much more challenging for the tools to successfully respond. Furthermore, synoptic assessments frequently

		incorporate open-ended questions or tasks that also require skills in critical thinking and analysis. These areas challenge AI, which may struggle to grapple with nuanced problems or demonstrate genuine understanding beyond rote memorisation.

Table 1: Assessment formats and their potential vulnerability to generative AI tools.

Mitigations to guard against the negative effects of generative AI are possible for all assessment types, but some require significantly more work to design and implement by academic staff than others.

Designing assessments to mitigate the influence of generative AI tools

Assessments that are designed to help mitigate the effects of generative AI will help to promote fairness and ensure that outcomes better reflect a student's own individual knowledge and skills. Assessment strategies that promote academic integrity more broadly (Holden, Norris, and Kuhlmeier, 2021) can be effective in reducing the ability of students to use generative AI tools within their assessments. Similarly, by academic members of staff developing an understanding of how generative AI tools behave when prompted, assessments can be modified at either a programme or module level to further limit its influence.

Assessment diversity

A good overall strategy is to include a diversity of assessment types within a programme, as assessments that are not text-based are currently less vulnerable to the effects of generative AI. More broadly, assessment diversity helps ensure fairness and inclusivity by accommodating different learning styles and strengths thereby allowing all students to demonstrate their individual knowledge, understanding and skills. Simulating real-world tasks helps to facilitate skills development and can make the learning experience appear more relevant and authentic to students (Villarroel et al., 2018). Multiple assessments that are lower stakes not only help reduce student assessment anxiety, but also allow students to receive more diverse feedback to help appraise their own learning performance.

There are many different formats of assessment (Habeshaw, Gibbs and Habeshaw, 1993) but assessment variety must be designed at a programme level. This is essential to ensure a coherent assessment structure for the programme, and that students have sufficient opportunities to successfully prepare for, and receive feedback on, the different types of assessment they will experience throughout their studies. Care does though need to be taken to ensure that increasing assessment diversity does not inadvertently result in increased summative assessment loads for students.

Synoptic assessment

Synoptic assessment (Constantinou, 2020), also described in Table 1, requires students to demonstrate their understanding of the relationships between the different aspects of their course; either across the programme as a whole, or between multiple modules. It extends beyond assessing only knowledge and understanding, the lower levels of Bloom's Taxonomy (Bloom, 1956), to assessing skills in application, analysis and evaluation. It therefore promotes deeper and more connected learning, whereby students develop a broader view of a subject, and helps better prepare them for the transition to the workplace where there is a need to draw upon knowledge from various sources to solve complex real-world problems.

Scaffold assessment tasks

Scaffolded assessments, introduced in Table 1, build upon each other over the course of a term or semester, culminating in a single summative grade for a larger and more complex piece of work that demonstrates how a student has achieved the intended learning outcomes. A scaffolded assessment involves the overall assessment being broken down into a linked series of smaller steps. It allows students to build and develop their skills incrementally, receive feedback at multiple stages to help them appraise their progress and improve their work, and better manage their assessment workload thereby reducing stress and anxiety. From the perspective of an academic member of staff, it allows familiarity with the work of individual students to be gained as the assessment progresses. Box 1 shows how a final year, and year-long, project in mathematics might be scaffolded with a mixture of

individual and group-based tasks. Each individual task may contribute a low stakes weighting towards the final assignment mark.

Learning outcome: Students are able to analyse the historical development of a mathematical topic area, evaluating the significance of individuals to the current collective knowledge of the field.
Final Assessment Task: A 15-page individual written report.

Task 1: Students identify a topic area and key individuals responsible for its development. A small group presentation is delivered on the contribution of a key mathematical figure and their connected relationships with other mathematicians.
Task 2: A list of key references to support the topic is developed along with a short video or audio-based summary of five key works.
Task 3: Students develop an outline plan or structure for their report which is presented to a small peer group. Students are required to provide feedback to each other on their plans and each individual student is required to reflect upon how they will modify their plan as a result.
Task 4: Students develop a defined section of their report, for example an introduction or background, which they submit to an academic member of staff for feedback.
Task 5: Students submit a draft of their report for peer and staff member feedback.

Box 1: Scaffolded assessment within a final year project in mathematics.

Assess the process of learning, not only the final output

Often, much of the assessment of whether a student has met the learning outcomes of a module is based upon a single piece of work, such as an examination or a project. Here the emphasis changes from assessing only the final output, to assessing the development of the student that takes place during the learning process. By evaluating how students engage with the process of learning, the strategies and techniques they use, and their reflections

upon doing so, it allows insight to be gained into how students think and how their skills and abilities to manage their own learning develop with time. As the emphasis is upon evidencing a continual process rather than a single final output, the assessment is much more resilient to the effects of generative AI tools which currently find it hard to provide evidence of personal growth over longer timescales. Examples of assessment approaches that might be used are shown in Box 2.

Notebooks: Requiring students to keep a notebook detailing the strategies, steps and approaches they are using to tackle a problem along with the challenges they are experiencing and what, and how, they are learning allows the process of inquiry and application to be assessed.

Reflective logs: Similarly, reflective logs, which might be written or presented in multimedia forms, assess the ability of students to reflect upon their learning and skills development throughout the process. They help students identify areas they find challenging, how they might address these challenges, and where they require further support.

Sketches and drawings: Sketches and drawings, which generative AI tools struggle to create reliably and consistently, can also be used to document learning. Examples include sketches of experimental set-ups, graphs, and mappings showing connections between ideas.

Portfolios: Discussed previously in Table 1, students could be asked to compile a portfolio of material alongside the development of a larger piece of work such as a project. For example, a portfolio might contain collections of academic papers, subsequent student analysis of them, and reflections ahead of, or immediately after, a supervision session.

Box 2: Approaches that allow assessment of the learning process

Change where assessment takes place and who is involved

Assessments that are particularly vulnerable to the effects of the inappropriate use of generative AI tools are typically unsupervised. One approach is therefore to incorporate more assessments that take place in-class, for example by having students develop a plan for

an assessment during a teaching session or working collaboratively with other students on a task. Students might be asked to review a case study prior to a teaching session and submit their questions in advance. They then either answer the questions or present on the case study during the session; this might take place individually or in smaller groups. Whilst this may not prevent students initially using generative AI tools, they will need to understand the material sufficiently well to present it and answer questions successfully. Research has shown that "the least acceptable forms of behaviour are those that disadvantage other students" (Ashworth, Bannister and Thorne, 1997, p. 198) and so group-based tasks can also be effective in reducing the motivations and opportunities for students to engage in the inappropriate use of generative AI tools, particularly when students are asked to assess their own, and their peers' contributions.

Add a local or disciplinary context

Although the ability of generative AI tools to access real-time information is improving, there remain limitations upon the level of the latest material that they can access. For example, many research papers are currently only accessible via an institutional subscription, and recent, and quite specific, institutional or disciplinary case studies are unlikely to have been widely used within their current training datasets. Creating assessments that are highly specific to a subject or institution can not only help enhance student motivation and engagement with the task but can also limit the ability of generative AI to be used. Students might be asked to develop specific course materials, examples or notes, or comment upon real or fictional case studies using a course-informed perspective. Assessments might draw upon previous modules they have studied, or follow-on from material discussed in teaching sessions, for example by encouraging engagement with a discussion board or the development of a wiki.

Increase the use of real-time dialogue and questioning

Oral assessments (Joughin, 2010) require students to respond in real-time to a series of verbal prompts and so their spontaneous nature not only makes them more resilient to the effects of generative AI, but also allows for a more genuine assessment of student knowledge, understanding and skills. However, care is needed because whilst they can aid

students with their skills development, many find them intimidating, and a challenge is ensuring fairness when multiple assessors are used. The use of oral assessments therefore requires appropriate scaffolding over the course of a programme, beginning in a formative way. But they can be introduced in informal ways. For example, a conversation following the development of a practical task may take place between a student and an academic member of staff, a teaching assistant, or if a clearly defined rubric is available, a fellow student. Whilst the *viva voce* associated with a PhD examination is a well-known example of an oral assessment, other approaches involve the OSCE (Observed Structured Clinical Examination), which is widely used in medical education because it allows students to practice and demonstrate clinical skills within a standardised medical scenario (Harden, 1988).

Target and reward higher order thinking

Assessment tasks involving the recall of knowledge or basic knowledge application that can create the perception of understanding, that is those at the lower levels of Bloom's taxonomy (Bloom, 1956), are highly vulnerable to the effects of generative AI. Tasks that assess higher order skills are much more resilient. In addition to assessment approaches included within Table 1, practical assessment tasks might include: (analyse) requiring students to participate in a debate based upon information contained within a series of conflicting research studies; (evaluate) asking students to review case studies or examples of work, which could be presented in a variety of formats, against defined criteria; or, (creating) developing a business plan or marketing strategy for an organisation based upon either a real dataset or a specifically designed case study. Whilst generative AI tools might offer some insight into the tasks at their lower levels, revising marking and grading schemes will ensure that where higher-level skills are successfully demonstrated they can be appropriately recognised and rewarded.

Be prescriptive

For assessment of the lower levels of Bloom's Taxonomy, which will typically be essential in the early stages of an undergraduate programme, being prescriptive with the format and structure assignments can offer some success in mitigating the effects of inappropriate generative AI use. For a written document, strict word limits and required section structures

might be specified within the assignment brief; similarly, the requirement to include user-generated images with appropriate legends, the use of a particular referencing format, or a specific style of writing and treatment of acronyms.

Be formative

Where assessment is high stakes, the use of generative AI tools by students can misrepresent the data relating to their academic performance and progress. However, the use of such tools may be far less problematic, or likely, for assessment formats that are either lower-stakes, or entirely formative, particularly when coupled with high-quality feedback that then supports future learning. The increased used of formative assessment components can therefore not only enhance student learning, but it allows information to be gathered about whether subsequent student performance in summative assessment components is consistent with this earlier work. The use of formative assessment provides an ideal opportunity to highlight to students the potential risks to the development of their disciplinary knowledge and skills if they become reliant upon the use of generative AI tools, or use them inappropriately, as part of their learning experience.

If you can't beat it – use it: using generative AI tools to enhance the assessment process

Using AI tools successfully, and ethically, will become an increasingly important skill within many future careers, and so it is important that students are given opportunities to develop this skill naturally within their programmes of study. Where its use within an assignment is permitted, the parameters surrounding how it can be used must be made clear to students, along with how its use within their work should subsequently be cited. If an assessment is summative, students should have first had opportunities to engage with its use in similar assessments in a more formative way. Students should also be informed of its limitations and the ethical issues associated with generative AI use including privacy and data considerations; potential for bias; inaccuracy and misrepresentation of information; ethics codes; plagiarism; sustainability; and, exploitation.

Table 2 provides examples of how the use of generative AI tools can be incorporated into assessments and used by students. This list will inevitably grow as academic staff becoming increasingly familiar with such tools and innovative within their teaching. Where generative AI is used to create resources or materials, it is important these are checked carefully prior to their release.

Ask students to review and grade generative AI outputs	For written outputs, students can critique the response provided by generative AI tools. This might involve assessing the accuracy of the output, its strengths and limitations, the expected or missing features, areas where further information or detail is needed, and the validity of any references. The required use of a grading scheme to assess the output will help students develop their own skills in academic writing.
Ask students to modify generative AI outputs	A natural extension is to ask students to modify the outputs from generative AI tools; where text-based tools are used this could be via track-changes, but annotated corrections could be made for mathematical or scientific disciplines. The fact that such tools are prone to errors and misconceptions, which can sometimes be quite subtle, presents an opportunity to help students learn and develop their understanding of topic areas, as well as enhancing their wider skills development.
Gain feedback prior to submission	Students can be encouraged to use generative AI to provide feedback on initial drafts of their work. They might then be asked to develop a plan of how they intend to respond to the AI generated feedback or to highlight where, and why, they disagree with it. This can form a useful exercise in helping students understand the nature of feedback and how to use it effectively to enhance their learning.
Creation of case studies or datasets	Generative AI offers the potential for academic staff to create more personalised learning resources quickly and at scale. For example, it is possible to create multiple case studies, all different but based upon a common theme or underpinning structure, which can then be used by individual students. Similarly, generative AI tools can be used to create individualised qualitative and quantitative datasets for subsequent student analysis.

Data analysis	Some generative AI tools offer powerful data analysis capabilities. Students might be encouraged to use such tools to perform an initial analysis and asked to explore how this compares with their own subsequent analysis. Where there are differences, these can then be explored in depth.
Research: Literature reviews	Generative AI tools might be used by students to analyse and/or synthesise one, or more research papers, case studies, or more extensive reports. Students could be asked to review the resulting output to explore whether the expected key ideas and themes are included. Students might cross reference the generative AI output with their own analysis/synthesis to identify similarities and differences, or use it to identify further literature sources to consider within their work.
Research: Developing and refining research questions	Following on from an analysis of key literature, students might be asked to use generative AI tools to identify and develop related research questions for their own projects and dissertations which they can subsequently refine. Such tools might be used to help create the structure and unique focus of their subsequent work or to identify key terms or ideas requiring further research.
Image creation	Generative AI tools can be used to add more creative or visual elements, such as audio or video, to a piece of work. Examples might include creating concept maps that define the relationships between disciplinary topics and ideas.
Improving academic writing	Generative AI tools can be used to provide proofreading assistance allowing students to correct any grammatical areas or shorten long sentences. It is important to ensure that any such usage is in line with relevant institutional codes of practice relating to third-party editorial assistance.
Developing glossaries	Students might be asked to use generative AI tools to assist in the first instance with developing glossaries of key subject specific terms or ideas. These might be refined through further individual research or peer discussion and aligned to relevant published literature or other information sources.
Computer coding	Generative AI can assist with computer coding, either the development or de-bugging of code, its documentation, or translating from one programming language to another. Instead of manually typing each line of code, a student might provide a generative AI tool with a description of what they want the code

	to do. This can help develop skills in precise prompting with students required to demonstrate how they have modified the AI generated code, and documented the output, to tackle a specific problem case.
Translating content to different forms	Students might be asked to use generative AI tools to develop an initial script for a video or podcast. Before recording, this could be refined to allow for a more subject-specific or localised context.
Generating alternative text for images	Alternative text conveys the content of an image in a non-visual manner via a textual description of what it presents. This is essential for accessibility and to aid individuals who might be exploring a body of text but who do not have the ability to visually experience images contained within it. Generative AI tools might be used to initially develop alternative text which students can then refine; this helps develop their ability to communicate complex information in a concise and precise manner.

Table 2: incorporating the use of generative AI tools into assessments.

When allowing, or encouraging, students to use generative AI tools within their assessments, either summative or formative, it is important to consider the issue of equity: that is all students should have equal access to equal tools. Subscription versions of the current mainstream generative AI tools perform, as expected, much better than their free-to-use counterparts. Academic staff should therefore require the use of the free versions of these technologies by all students, or better, mandate the use of an institutionally approved generative AI tool which will often be implemented in a more information-secure manner. This might be checked by requiring the inclusion of screenshots demonstrating the prompts used and the resulting response of the AI tool.

The non-permitted use of generative AI tools

Where it is important to ensure that AI tools are not used within an assessment, additional mechanisms are required to protect the integrity of the assessment process from the unauthorised use of generative AI technologies. Whilst examinations will continue to have an important role within assessment, an increased move to proctored unseen examinations should not be seen as a proportionate part of the response, and as such their use needs to

be appropriately considered relative to the learning outcomes being assessed. As French, Dickerson and Mulder (2023) note, "there is substantial evidence that examinations cause elevated distress and anxiety...the proven adverse effects of examinations on student mental health and wellbeing is concerning, as is the negative impact of examination anxiety on student motivation."

An alternative approach might be to conduct assessments on campus in more informal settings, but additionally utilise technologies that restrict access to generative AI tools. However, using such 'blockers' is unlikely to be effective. Students can often bypass their use, but more significantly, generative AI tools are becoming increasingly embedded within a range of common software and online technologies, some of which may be necessary as part of the assessment process.

It is appropriate to mention here the use of generative AI detection tools, of which many are now advertised, as these are unlikely to ever prove reliable at detecting AI-generated content. This reinforces the need to think carefully about assessment design. AI detection tools work by using very large data sets, collected from a variety of sources, to predict the likelihood of certain words or phrases within a particular passage of text. Many of the detection tools are based upon machine learning classifiers, algorithms that automate the ordering or categorisation of data into one or more 'classes'. Perhaps the most common example is an email classifier that scans emails and filters them into the classes of 'spam' and 'not spam'. Algorithms are used to determine patterns in the text, with the more highly predictable the next word relative to the previous, the more likely the detection tool to determine that it is AI generated. By considering an entire body of text and the patterns within it, the detection tool reaches a conclusion based upon the content. However, in a real-world situation, this approach does not always work well with the tools giving rise to both false-positives (classifying human-authored text as machine-generated) and false-negatives (classifying machine-generated text as human-authored).

There are a growing number of studies that explore the effectiveness of AI detection tools and yield similar conclusions that these are not reliable in determining whether text-based content is AI-generated. Elkhatat, Elsaid and Elmeer (2023, p. 1) found *that "*AI detection

tools were more accurate in identifying content generated by GPT 3.5 than GPT 4. However, when applied to human-written control responses, the tools exhibited inconsistencies, producing false positives and uncertain classifications*"*. Weber-Wulff *et al.* (2023, p. 1), whose work also considered the commercial AI-generated text detection systems of Turnitin and PlagiarismCheck, concluded that "the available detection tools are neither accurate nor reliable and have a main bias towards classifying the output as human-written rather than detecting AI-generated text" and furthermore "content obfuscation techniques significantly worsen the performance of tools". There is also increasing evidence that "GPT detectors frequently misclassify non-native English writing as AI generated, raising concerns about fairness and robustness" (Liang *et al.*, 2023, p.1).

The importance of good assessment design

Although strategies for mitigating, or embracing, the developments of generative AI tools upon assessments have been considered here, it is important to ensure that these are more broadly underpinned by principles of good assessment design. These include the need to ensure a holistic view of assessment across a programme rather than at a module level and that assessment strategies should be carefully designed to be inclusive for all learners. Varied assessment is naturally more inclusive: different methods of assessment may advantage or disadvantage different students or groups of students, and so offering variety allows every student the best possible chance to demonstrate their knowledge, skills and understanding. If implementing assessment types that might initially appear unfamiliar to students, they should have opportunities to first engage with them in a formative way; this includes tasks involving groupwork.

Summative assessment has a role in determining whether students have met the learning outcomes of their programmes. Yet whilst such assessment is important for helping students understand what they have learned, formative assessments allow students to demonstrate the learning journey that they are on within a risk-free environment. Where effective feedback is provided in a timely manner, students can use this to appraise and enhance their learning gain. It is therefore vital that when modifying assessments in response to generative AI tools, this balance between assessment **of** learning, versus assessment **for** learning is at

the forefront of considerations, and that summative assessment loads are reviewed to ensure they are proportionate and balanced.

Acknowledgements

As perhaps might be expected given its title, generative AI tools have been used within the development of this work as part of an experiment. Their primary use, with some success, has been to support more efficient information retrieval as part of the research process.

References

Ashworth, P., Bannister, P., Thorne, P., & Students on the Qualitative Research Methods Course Unit. (1997) 'Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment', *Studies in Higher Education*, 22(2), pp. 187-203. https://doi.org/10.1080/03075079712331381034.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., and Krathwohl, D.R. (eds) (1956) *Taxonomy of educational objectives: the classification of educational goals; Handbook I: Cognitive domain*. New York: David McKay.
Brown, G.A., Bull, J., and Pendlebury, M. (2013) *Assessing student learning in higher education*. London: Routledge.

Buckley, A. (2023) 'Are we answering the question that has been set? Exploring the gap between research and practice around examinations in higher education', *Studies in Higher Education*, pp. 1-17. https://doi.org/10.1080/03075079.2023.2283784.

Constantinou, F. (2020) 'What is synoptic assessment? Defining and operationalising an as yet non-mainstream assessment concept', *Assessment in Education: Principles, Policy & Practice*, 27(6), pp. 670-686. https://doi.org/10.1080/0969594X.2020.1841734

Elkhatat, A.M., Elsaid, K., and Almeer, S. (2023) 'Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text', *International Journal of Education Integrity,* 19 (17). https://doi.org/10.1007/s40979-023-00140-5.

French, S., Dickerson, A., and Mulder, R.A. (2023) 'A review of the benefits and drawbacks of high-stakes final examinations in higher education', *Higher Education*, pp. 1-26. https://link.springer.com/article/10.1007/s10734-023-01148-z.

Habeshaw, S., Gibbs, G., and Habeshaw, T. (1993) *53 interesting ways to assess your students*. Trowbridge: Cromwell Press Ltd.

Harden, R.M. (1988) 'What is an OSCE?', *Medical teacher*, 10(1), pp. 19-22. https://doi.org/10.3109/01421598809019321.

Holden, O.L., Norris, M.E., and Kuhlmeier, V.A. (2021, July) 'Academic integrity in online assessment: A research review, *Frontiers in Education*, 6. https://doi.org/10.3389/feduc.2021.639814.

Joughin, G. (2010) *A short guide to oral assessment*. Leeds: Leeds Met Press in association with University of Wollongong.

Liang, W., Yuksekgonul, M., Mao, Y., et al. (2023) 'GPT detectors are biased against non-native English writers', *Patterns*, 4 (7). https://doi.org/10.1016/j.patter.2023.100779.

Liu, D. L. [Danny Liu]. (2023, April 27). Prompt engineering for educators – making generative AI work for you [Teaching@Sydney]. Available at: https://teaching.sydney.edu.au/prompt-engineering-for-educators-making-generative-ai-work-for-you/. Accessed 2 July 2024.

University of Birmingham. (2024) *Generative Artificial Intelligence and its role within teaching, learning and assessment*. Available at: https://www.birmingham.ac.uk/schools/higher-education-futures-institute/generative-artificial-intelligence.aspx. Accessed 2 July 2024.

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., and Herrera-Seda, C. (2018) 'Authentic assessment: creating a blueprint for course design', *Assessment & Evaluation in Higher Education*, 43(5), pp. 840-854. https://doi.org/10.1080/02602938.2017.1412396.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. (2023), 'Testing of detection tools for AI-generated text', *International Journal for Educational Integrity*, 19(1). https://doi.org/10.1007/s40979-023-00146-z.

To cite this article:

Grove, M.J. (2024). Generative AI technologies and their role within assessment design. Education in Practice. Vol.5, Issue 1, pp 15 – 35.