Online Maths for Social Science Assessment Design in the Age of Generative AI

Soobedar de Villeneuve, Zeenat

King's Foundation, King's College London

1 November 2024

Online Maths for Social Science Assessment Design in the Age of Generative AI

Zeenat Soobedar de Villeneuve [a]*

[a]*King's Foundation, King's College London, London, UK*

King's Foundation, Centre for International Education and Languages (CIEL), King's College London, Strand, London WC2R 2LS, UK.

Email: zeenat.1.soobedar_de_villeneuve@kcl.ac.uk

ORCiD: 0009-0003-4924-230X

Abstract

In the fast-changing educational world, Generative AI (GenAI) has brought about big changes, especially in online formative and summative assessments. Universities are concerned about the unethical GenAI use, compromising academic integrity. This study proposes and evaluates strategies to design questions in maths for social science education that are challenging for ChatGPT-3.5 to solve. Drawing on Bloom's Taxonomy, a trend analysis of academic performances and focus group discussions, it proposes a transformed approach to assessment design: the SHARP (Strategic, Holistic, Adaptive, Reflective, Process) assessment cycle. This framework is iterative and integrates real-time feedback to ensure inclusivity and transparency, stemming from a Reflect-Rewrite-Retest-Review redesign approach, focusing on higher-order cognitive questions. A quantitative analysis between 2020 and 2024 reveals a significant increase in higher-order level questions (e.g. from 29% to 84% in a test) and a significant but not drastic drop in academic performance. The effectiveness of ChatGPT-challenging designs is corroborated by focus group discussions, highlighting the need for a balance between student accessibility and academic rigour. This study contributes to the literature by providing unique empirical evidence on the validity of the strategies and offering actionable steps for educators, policymakers and institutions to maintain academic integrity in maths for social science education.

Keywords: ChatGPT; academic integrity; formative and summative assessments; maths questions; assessment design

1. Introduction

The rise of remote learning, particularly due to the Covid-19 pandemic, has accelerated the transitioning towards online assessments. However, this shift has also increased threats to academic integrity (St-Onge et al. 2022). Assessments, as a crucial component of online learning, serve to evaluate students' progress and comprehension. However, online assessments can be vulnerable to various forms of academic misconduct (Clarke et al. 2023), including plagiarism, unauthorized use of resources, and multiple attempts at assessments. The pandemic further exacerbated this issue, correlating with a significant rise in cases of academic misconduct (Roe et al. 2024; Henderson et al. 2022; Lancaster and Cotarlan 2021) and altering perceptions of academic dishonesty among students and staff (Amzalag et al. 2021; Reedy et al. 2021). In this evolving educational landscape, Generative AI (GenAI) poses both opportunities and challenges. While it offers interactive capabilities that can transform learning by enhancing critical thinking skills (Bitzenbauer 2023), assisting in designing questions tailored to student needs (Su & Yang 2023), it also facilitates academic dishonesty by enabling students to automate or receive assistance in completing assignments. GenAI's can compromise assessments' authenticity (Eke 2023) with their ability to generate plausible answers, particularly to those that focus on recall or application of knowledge, in subjects like mathematics for economics or finance.

Traditional assessments which usually focus on lower order thinking skills are now more vulnerable to GenAI manipulation. Those questions are typically easy to handle by GenAI due to their predictable structure and straightforward content. Nikolic et al. (2023) highlights that assessments in engineering education on replicable problem-solving processes are particularly at risk. Essays or open-ended questions can easily be answered raising academic integrity and originality issues ((Van Dis et al. 2023). The use of GenAI in group projects may exacerbate "free riding" problems where some team members rely on those tools to produce work, hence reducing collaboration and engagement in teamwork. Dissertations and individual projects also face significant challenges given their emphasis on independent and original work. ChatGPT-like tools can generate text which lacks depth and references which are non-existent. Using GenAI to write dissertations or research abstracts undermines the analytical and critical thinking skills (Else 2023). For online tests, correct answers can be easily generated by ChatGPT undermining subject mastery and analytical skills of the students. These assessments are often designed to test students' understanding and ability to apply concepts to specific scenarios.

However, online tests have distinct advantages, including ease of grading, instant feedback (Eke 2023), and benefits for neurodiverse students who may feel less constrained by traditional, in-class testing environments (Dunne and Lee 2022). Moreover, timely feedback has been shown to positively impact student learning (Gikandi et al. 2011). So, if online assessments were here to stay, examiners should ensure rigorous assessment design to maintain their aim and integrity ((Nikolic et al. 2023). Institutional guidelines from organisations like (UCL Assessment Working Group 2020) for designing online exams recommend the inclusion of complex questions that require critical thinking, as well as educating students about the ethical use of GenAI. Holden, Norris, & Kuhlmeier (2021) suggest other strategies to reduce cheating in online quizzes, like carrying out exams at the same time within a short time window to limit the risk of students sharing test questions and answers. Other measures could include disabling the copy/paste functionality in assessment software, allowing questions to appear one at a time for a limited time without access to previous tasks once completed, randomising test questions and/or response options, creating multiple test versions when collaborating and sharing concerns exist. Different types of assessments for the module/course could place heavier weights on assessment where cheating is less likely. Statements of academic integrity should be clear with respect to GenAI use, plagiarism and collusion to help with students' expectations. To minimise plagiarism, Vellanki et al. (2023) further propose using a 'safe exam browser' option and more subjective-type (short answer) rather than objective-type (fill in the blanks) questions.

However, un-proctored online tests remain prone to some sort of academic misconduct, even when exam browsers options are available. This therefore calls for educators to rethink their assessment strategies and find innovative ways to ensure academic rigour, if not the use of GenAI in academic dishonesty is likely to escalate (Liu and Bridgeman 2023). When GenAI tools became more popular in early 2023, some universities banned it outright while others addressed adjusting assessments to permit its usage in a more ethical manner (Lye and Lim 2024). As GenAI is likely to stay, evolving and becoming increasingly influential in students' learning curve over time, it is important that assessments be redesigned to adapt to the changing educational landscape. Assessment designs should engage students with specific tasks that require critical thinking which cannot be easily replicated by Large Language Models like ChatGPT (Seo et al. 2021; Rasul et al. 2023). Research shows that assessments emphasising higher-order cognitive skills – such as analysis, evaluation, and creation – are less vulnerable to GenAI-generated responses. These skills are central to Bloom's revised taxonomy of educational objectives by Anderson and Krathwohl (2001). For example, assessments can take the form of open-ended questions and real-life applications.

The aim of this study is threefold:

1. Develop innovative strategies to design online maths for social science questions which challenges ChatGPT.
2. Quantitatively validate the effectiveness of the innovative strategies, mitigating the unethical use of GenAI in education.

3. Create a novel, iterative assessment cycle that reimagines assessment design by incorporating real-time student feedback, inclusivity and transparency in assessment design.

This paper breaks new ground in proposing strategies to create maths for social science questions that requires deeper reasoning beyond Gen AI capabilities using Bloom's taxonomy (Bloom et al. 1956) which has long been used in education to categorize learning objectives and create assessments but its use to counter GenAI-generated responses is especially novel. The study uses the analysis, evaluation and creation higher-order cognitive processes that Bloom identified as tools to lessen the efficacy of ChatGPT-3.5) in responding to test questions. This represents a substantial shift from conventional assessment design which frequently emphasizes lower-order cognitive abilities like recall and remember – domains in which GenAI is particularly strong. We draw upon additional theoretical models such as Cognitive Load Theory (Sweller et al. 2011), Constructivist Learning Theory (influenced by Piaget (1976), and Vygotsky (1980), Metacognition (Flavell 1979), and Critical Thinking Frameworks (Paul and Elder 2013), all of which emphasize promoting independent thought and problem-solving abilities. Designing questions that require learners to evaluate arguments, solve complex problems, or make decisions based on evidence can be particularly challenging for GenAI, which might lack the nuanced judgment required for critical thinking (Paul and Elder 2013). This approach ensures that questions engage learners at various levels of cognitive complexity (Bloom et al. 1956), The paper proposes and uses the 4Rs strategies (Reflect, Rewrite, Retest, Review) to ensure that online test questions are not easily solved by ChatGPT-3.5. This process is iterative, starting with a reflection on pre-ChatGPT questions that meet learning outcomes, followed by the redrafting of questions to target higher order thinking skills, retesting them using GenAI, and reviewing the results. If ChatGPT successfully answers the questions, the process restarts. Such a novel approach to assessment design ensures that learning outcomes are achieved without compromising the clarity and meaning of the questions.

The paper further validates the newly designed strategies by analysing student grade trends over the past four years. Existing studies (Sallam et al. 2023; (Tan et al. 2023) primarily propose theoretical solutions without empirical evidence and real-world settings performance evaluation. By monitoring student performance and examining the direct effects of question design modifications on GenAI's response capabilities, the study provides a unique empirical examination of the efficacy of questions that challenges GenAI. This study also extends the discussion on the ethical use of GenAI by exploring the impact of higher-order level questions on students' engagement with assessments and learning journey. Through focus groups, the study raises the importance of incorporating student voice in assessment design to balance rigour, inclusivity and academic integrity.

In conclusion, this study adds to the international knowledge base with its analysis of the effectiveness of strategies used to develop more rigorous assessments in the age of GenAI. Unlike the current literature which highlights theoretical or anecdotal solutions, this research quantitatively tests and evaluates the innovative question designs in a real-world educational setting. It further proposes a novel assessment framework, the SHARP assessment cycle, which integrates real-time student feedback and inclusivity. As such, the research strengthens the weak link between theoretical underpinnings and real-world applications to uphold academic integrity.

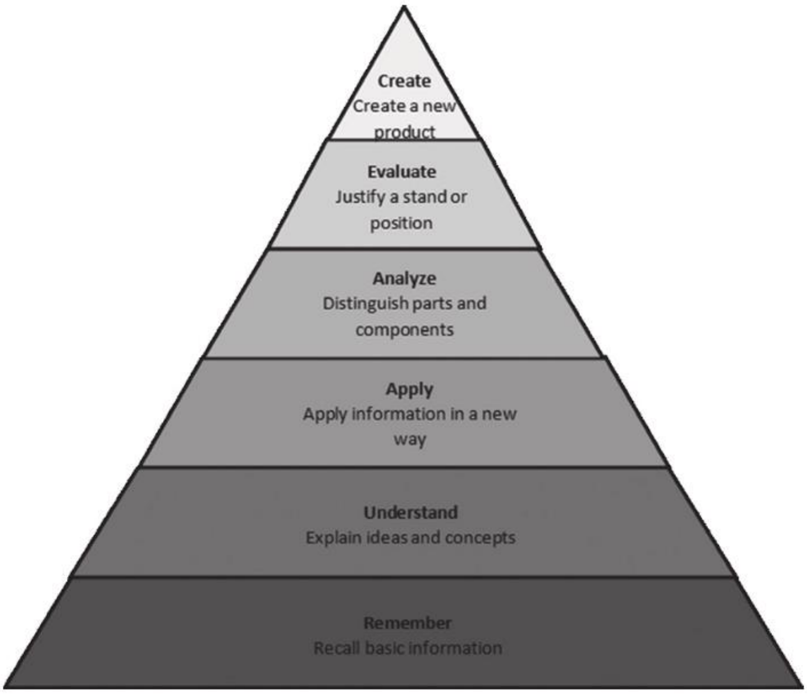
The paper is structured as follows: Section 2 outlines the methodology detailing the 4Rs approach followed by a more detailed explanation of the Rewriting and Retesting phases. Section 3 presents findings followed by a discussion of results in Section 4, where the SHARP assessment cycle is developed. Section 5 concludes.

2. Methodology

2.1 the Revised Bloom's Taxonomy of Educational Objectives

Fig. 1 reports the guiding framework of Anderson and Krathwohl's revised Bloom's Taxonomy of Cognitive Domain ((Phillips et al. 2019) underpinning this paper. Drawing further upon the Cognitive Load Theory and Constructivist Learning Theory established educational frameworks, the approach to question design fosters independent and critical thinking. The application of those theories ensure that questions focus on higher-order thinking skills, which GenAI find more challenging to respond, thereby addressing a gap in the current literature. The study applies this approach to designing maths for social science questions.

Fig. 1 The revised Bloom's taxonomy of educational objectives



Source: Phillips, Briggs, and Jensen (2019)

Recall (or Remember) Questions. These require retrieving definitions, facts, or formulas. Since these questions rely on memory, they are typically the easiest for both students and ChatGPT. Examples include questions that ask for definitions of terms or the recall of a specific formula to solve a problem (e.g., finding the mean from raw data).

Understanding Questions. These assess students' ability to explain, interpret and compare concepts. Students are required to interpret information and draw relations between concepts, such as explaining the connections between demand and supply. Although ChatGPT can work these tasks effectively, it lacks depth in interpretation.

Application Questions. These relate to solving problems, performing calculations or analysing specific scenarios. For instance, a student is asked to calculate the impact of a subsidy on demand and/or supply curves. Even though ChatGPT can handle standard applications, it may find it harder when problems are unique to certain contexts or subject to specific economic constraints (e.g. total output cannot be negative).

Analysis Questions. These test students' ability to break down complex information into sub parts and examine any relationships between them. This may involve a trend analysis or critical evaluation of datasets. For example, students are asked to choose the correct formula to calculate the real average salary from a list of numbers. In this case, they are expected to dissect and scrutinize various components of the formulae. This task challenges ChatGPT as it may struggle in contextualising relationships among known values.

Evaluation Questions. These require forming a judgment, an argument or an assessment based on certain criteria. For instance, students are asked to evaluate whether an economic model is valid based on certain assumptions. Evaluation tasks are more complex and required higher order level thinking skills enabling students to assess which options they have and justify their conclusions. This process demands subjective reasoning and judgment, criteria with which ChatGPT struggle with.

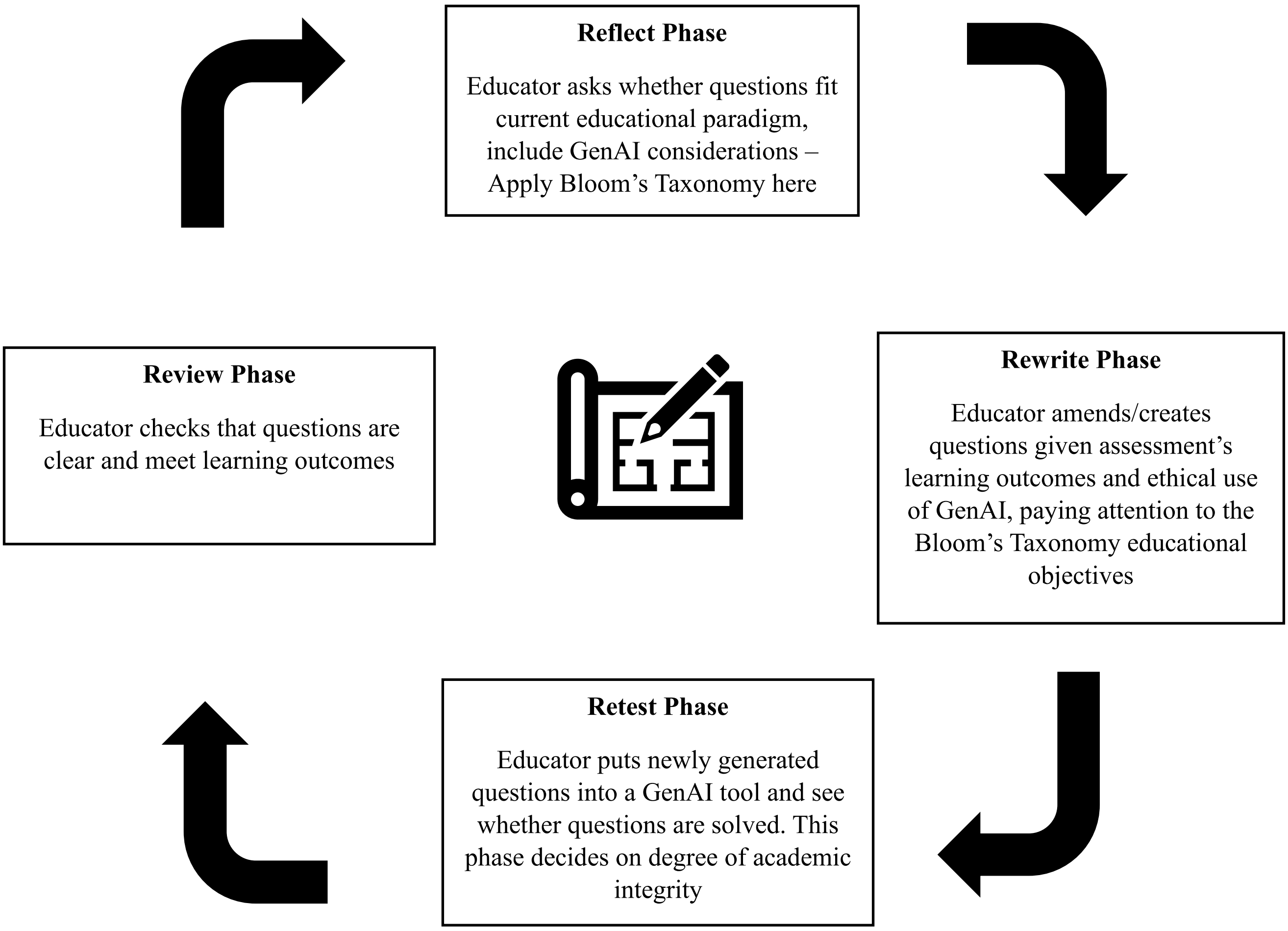
Creation Questions. These involve students to retrieve and combine information from various areas to generate new ideas and/or solutions. Creation tasks, which fall in the top category of Bloom's taxonomy, are particularly challenging for ChatGPT. For example, using Bayes' theorem (a statistics topic) with missing information on certain probabilities, students need to work backwards to calculate the probability of an event happening before applying the theorem to answer the question. This process requires innovative thinking than merely applying formulae, which GenAI systems are incapable of doing without specific prompts.

2.2 The Process of Assessment Design: The 4Rs Approach

The process used in designing assessment higher-order level questions includes four phases: the reflect, rewrite, retest and review phases. The process begins with the use of a pre-existing set of questions or the creation of new ones - the reflect phase. The educator goes through pre-existing questions (or if creating questions), reflecting on how they relate to Bloom's Taxonomy educational objectives. Questions are then input in one of the GenAI tools (ChatGPT, Gemini, Copilot) and responses are recorded. If GenAI answers correctly, it is then asked how it may struggle with the question. The question is subsequently rewritten, retested and reviewed, again in line with Bloom's Taxonomy. The process is repeated until GenAI can no longer provide a suitable answer[1] as shown in Fig. 2.

[1] Nikolic et al. (2023) used a similar method although their approach was not as explicitly detailed in steps as presented in this paper.

Fig. 2 the 4Rs approach



2.2.1 Reflect Phase: The Theoretical Framework Relating to Cognitive Skills

Based on Bloom’s Taxonomy and its revised versions (Anderson and Krathwohl 2001), this stage seeks to categorise questions in the six key levels of cognitive skills: recall, understand, apply, analyse, evaluate, and create. These important skills are required in mathematics and applied sciences to move students beyond rote memorization and engage with problems on deeper levels (problem-solving, analytical and critical analyses (Clements and Battista 1990).

2.2.2 Rewrite and Retest Phases: Designing GenAI-resistant Questions by Question Types

The next stage to mitigate ChatGPT's strengths in solving straightforward problems, involves the testing (Retest) phase where ChatGPT’s responses to the original questions are analysed. Following this, a rewriting phase as the name suggests consists of steps to modify the original questions to challenge ChatGPT’s capabilities more effectively. Below are some examples.

Recall Question Example (Topic: Raw Data – Measures)

Retest stage.

PreChatGPT Question and Human Answer

In **Fig.** 3 panel a, students were asked to calculate the range directly from a dataset.

Fig. 3 panel a: "Recall" question and human answer - finding range from raw data

The data set below shows the number of items purchased by 10 shoppers at Selfridges (London) in a year: $12, 23, 25, 26, 54, 92, 27, 52, 36, 64$.

Compute the range, the upper quartile, the median and the quartile range. Give your answers to the nearest whole number.

Enter the values below:

a) The range is 80

ChatGPT Answer

Correct, as expected for recall-type questions where formulas are explicitly provided. **Fig.** 3 panel b shows the steps ChatGPT uses to solve this question.

Fig. 3 panel b: correct answer from ChatGPT with workings

To compute the range, upper quartile, median, and quartile range, let's first sort the data and then apply the relevant formulas.

The data set: $12, 23, 25, 26, 54, 92, 27, 52, 36, 64$

Step 1: Sort the data in ascending order.

Sorted data set: $12, 23, 25, 26, 27, 36, 52, 54, 64, 92$

Step 2: Compute the range.

The range is the difference between the maximum and minimum values.

$$\text{Range} = 92 - 12 = 80$$

Rewrite stage

ChatGPT Proof Question and Human Answer

In **Fig.** 4 Panel A, the above question was redesigned as a creation-type problem where students were expected to recognise that there are missing values from the dataset, making the range impossible to find.

Fig. 4 panel a: "Creation" question and human answer - finding range from raw data

a. A restaurant chain asked managers of its 10 branches to report how many customers were served on a particular day. The following data set represents the numbers reported by 9 managers: $25, 23, 12, 26, 54, 92, 51, 36, 64$.

What is the range of the number of customers served by the restaurant chain that day?

cannot be determined

With drop down menu:

80
39
40
52
cannot be determined

ChatGPT Answer

Incorrect. The change of question type from a recall to creation task increased the difficulty level significantly. This time students needed to recognise the missing value and determine how this affected the solution, requiring higher order thinking to creatively interpret the incomplete information and infer possible outcomes rather than recall or apply formulas. **Fig.** 4 panel b proves that ChatGPT does not have those skills.

Fig. 4 panel b: incorrect answer from ChatGPT

To determine the range, we subtract the smallest number from the largest number in the data set.

Given data: $25, 23, 12, 26, 54, 92, 51, 36, 64$

Step 1: Identify the smallest and largest numbers.

- Smallest number: 12
- Largest number: 92

Step 2: Calculate the range.

$$\text{Range} = \text{Largest number} - \text{Smallest number} = 92 - 12 = 80$$

So, the range is 80. The correct answer is **80**.

Understanding and Analysis Questions Examples (Topic: Quadratic and Graphs)

Retest stage

PreChatGPT Question and Human Answer

In **Fig.** 5 panel a, students were asked to interpret a quadratic function's graph and answer understanding (part b) and analysis-type (parts c and d) questions, such as finding and interpreting the vertex.

Fig. 5 panel a: "Understanding" and "Analysis" questions and human answers – interpreting intercepts and vertices

Sketch the profit function $y = -45 - 6q + 3q^2$, where q is the total output. Then use your graph to answer the questions below.

Enter your answers into the boxes to complete each statement:

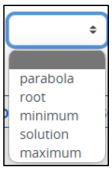
a) The curve crosses the q axis at [1] point/s.

b) The highest value of q for which profit is equal to zero is [5]

c) The curve has a vertex when $q =$ [1] and $y =$ [-48]

d) This vertex is a [minimum]

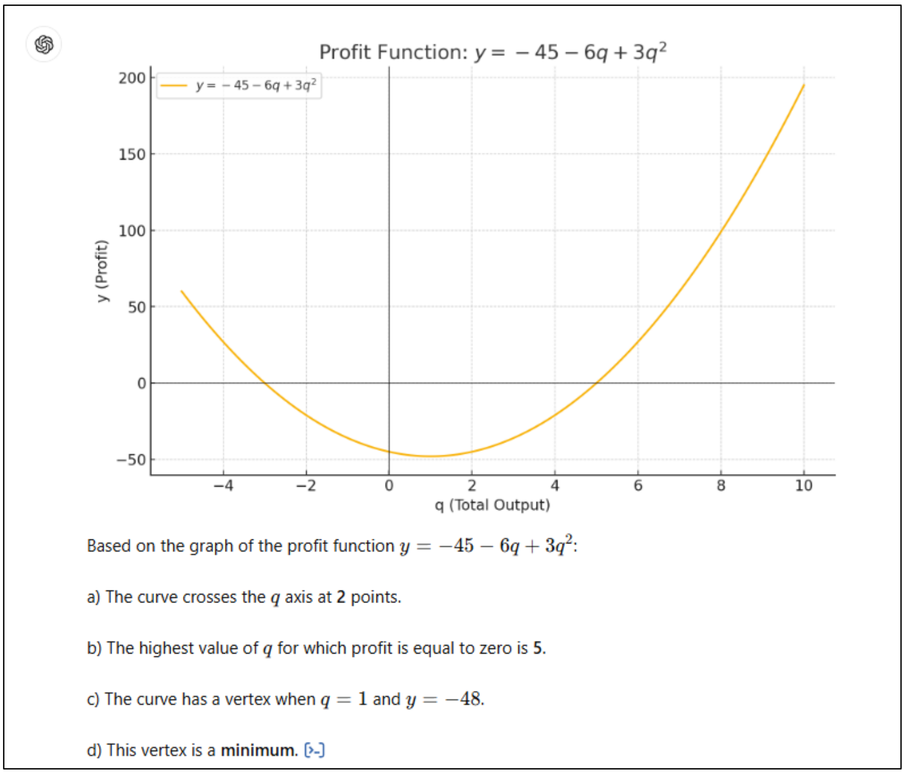
With part (a) with 0, 1 or 2 dropdown options and (d) with



ChatGPT Answer

Parts (b)-(d) Correct, as expected for understanding and certain analysis-type questions. (b) requires students to remember the relationship between 2 variables (quantity and profit). (c) and (d) are analysis-type questions because students need to break down the mathematical properties of the curve (key to analysis-level thinking) and find out the specific variables needed for the vertex in addition to determining how changing quantity and profit would influence the shape and position of the curve. A graph, as shows in **Fig.** 5 panel b below, was also generated, unexpected based on prior ChatGPT graphical limitations.

Fig. 5 panel b: correct answers for (b)-(d) from ChatGPT

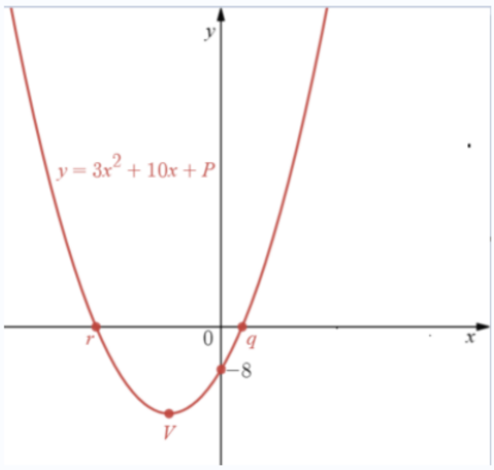


Rewrite stage

ChatGPT Proof Question and Human Answer

In **Fig.** 6 panel a, the above question was redesigned as a graph-type task where students were required to fill in missing components. Graph-type questions can be used as a last resort to make questions more ChatGPT-resistant.

Fig. 6 panel a: “Graph” question and human answer - finding missing information about intercepts and vertices



These questions refer to the sketch graph shown above.

Answer each question by typing a **number** into the box.

Each answer should be given as a whole number or a decimal rounded to 2 decimal places.

a. State the value of P in the equation of the curve. -8

b. State the value of r. -4

c. State the value of q. 0.67

d. Find the horizontal coordinate of point V. -1.67

e. If x is intended to represent the weekly number of sofas produced in a workshop and y the weekly profit (in £) the workshop expects, which of these restrictions must be applied to the graph shown?

x≥0

ChatGPT Answer

Incorrect. ChatGPT couldn't read PDFs or images. These questions required visual interpretation and more human insights. Students were forced to engage more actively with materials and focus on deeper conceptual understanding. However, this strategy could be circumvented (by smarter students) through the correct prompts. **Fig.** 6 Panel b shows how part (a) can be correctly answered with the correct prompt. As of October 2024, ChatGPT's ability to upload images and documents led to solving the question correctly. However, a slight adjustment can generate GenAI-resistant questions without a graph (see **Fig**. 6 panel c).

Fig. 6 panel b: prompt that can be used to answer part (a) correctly

If there is a graph with equation y=3x^2+10x+P. what is the value of P is the curve cuts the x-axis at 2 places and y-axis at -8

Given the equation $y = 3x^2 + 10x + P$, we need to find the value of P under the conditions:

1. The curve cuts the x-axis at two places, meaning the quadratic equation has two real and distinct roots.
2. The curve cuts the y-axis at $y = -8$.

Step 1: Analyze the intersection with the y-axis

The curve cuts the y-axis when $x = 0$.

Substitute $x = 0$ into the equation:

$$y = 3(0)^2 + 10(0) + P = P$$

Given that $y = -8$ at this point:

$$P = -8$$

Fig. 6 panel c: reformatted **Graphs** (Topic: Quadratic and Graphs) following ChatGPT's possibility to upload PDF's

Sketch the profit function $y = -45 - 6q + 3q^2$

a) The curve crosses the q axis at ... point/s.

b) The lowest value of q for which profit is equal to zero is...

c) Assume that the curve is vertically scaled by 2. Find the vertex coordinates (...,...).

d) This vertex is a ...

Human Answer

Sketch the profit function $y = -45 - 6q + 3q^2$

a) The curve crosses the q axis at **1** point/s.

b) The lowest value of q for which profit is equal to zero is **5.**

c) Assume that the curve is vertically scaled by 2. Find the vertex coordinates (**1,-48**).

d) This vertex is a **minimum**.

ChatGPT Answer: Incorrect answers for (a) and (b) and partially for (c)

Fig. 3: Incorrect Answer by ChatGPT

I can't view the contents of the image directly, but I can help you with your questions based on the function $y = -45 - 6q + 3q^2$.

Here's a breakdown of the points you asked about:

1. **The curve crosses the q-axis at ... point/s:**
   - We previously established that the function crosses the q-axis at two points: $q = 5$ and $q = -3$. So, the curve crosses the q-axis at 2 points.
2. **The lowest value of q for which profit is equal to zero is ...:**
   - The lowest value of q where the profit is zero is $q = -3$, as derived from the roots of the equation.
3. **Assuming the curve is vertically scaled by 2, find the vertex coordinates (..., ...):**
   - After scaling, the new function becomes $y = 2(-45 - 6q + 3q^2) = -90 - 12q + 6q^2$.
   - The vertex for this new function was calculated to be at $(1, -96)$.
4. **This vertex is a ...:**
   - Since the parabola opens upwards (as the coefficient of q^2 is positive), this vertex represents a **minimum** point.

Application Question Example (Topic: Supply & Demand Linear Model)

A3.1 Retest stage

PreChatGPT Question and Human Answer

In **Fig.** 7 Panel a, students were asked a basic application question on the demand and supply model, that is to find equilibrium points.

Fig. 7 panel a: "Application" question and human answer – solving for a value in a supply equation

In this question the wage in £ per hour is represented by W and the quantity of labour by L :

Answer each question by typing a number into the box:

a) In the labour supply function $W = 4L + 52$, find L when $W = 60$.

$L =$ 2

ChatGPT Answer

Correct, as expected for a straightforward application-type questions.

Fig. 6 panel b: correct answer from ChatGPT with workings

To find L when $W = 60$ in the labour supply function $W = 4L + 52$, you need to solve the equation for L.

1. Substitute $W = 60$ into the equation:

$$60 = 4L + 52$$

2. Subtract 52 from both sides to isolate $4L$:

$$60 - 52 = 4L$$

$$8 = 4L$$

3. Divide both sides by 4 to solve for L:

$$L = \frac{8}{4}$$

$$L = 2$$

So, $L = 2$ when $W = 60$.

Rewrite stage

ChatGPT Proof Question and Human Answer

In **Fig.** 8 Panel a, the above question was rephrased as an evaluation-type problem. In here, students were required to use their judgement and evaluate the feasibility of different scenarios, based on economic constraints.

Fig. 8 panel a: "Evaluation" question and human answer - finding unknown pay level given work hours

A painter works in a partnership with a carpenter and an electrician.

Each partner's total weekly income (in £) is represented by Y and is made up of a fixed weekly payment, plus a variable payment based on an hourly wage (in £), represented by H, and the number of hours worked this week.

The painter's fixed weekly income is £32, the carpenter's is £40, and the electrician's is £35.

Last week the three partners worked a total of 10 hours between them: the painter worked for 2 hours, the carpenter for 4 hours, and the remaining hours were completed by the electrician.

The equation that represents the total income of the electrician's salary is of the form $Y = aH + b$.

a. Find a and b. Answer each question by typing a number into the box:

$a =$ 4

$b =$ 35

ChatGPT Answer

Detailed response but Incorrect for part (a). ChatGPT gets that wrong because the question requires students to analyse the information about the partnership and how the hours worked by each partner relates with income. Students need to assess the relationship between income and hours worked in a specific context and justify the rationale for the choice of a and b based on given information and assumptions. Synthesizing information and making judgements are characteristic of the higher-order evaluation thinking skills of Bloom's taxonomy.

Fig. 8 panel b: incorrect answer from ChatGPT although lengthy workings

4. **Identify a and b:**

   In this context, b represents the fixed weekly payment, which is £35. Thus:

   $$b = 35$$

   a represents the variable part of the income based on the hourly wage. Since we do not have any information about the hourly rate, it is not specified in the problem, but it is assumed that a is the rate at which income varies with hours worked.

To summarize:

- a (the variable payment per hour) is not explicitly provided, so it remains as a in the general form.
- b is the fixed weekly income, which is 35.

Therefore:

- a: (It remains unspecified in the problem)

2.2.3 Review Phase: Evaluating Learning Outcomes

Throughout the whole process, the primary objective is to ensure that the revised questions remain aligned with the module's intended learning outcomes and retain their clarity and accessibility to students while introducing some level of complexity that challenged GenAI capabilities. In addition to comparing final answers, the review phase also includes the analysis of ChatGPT's reasoning process, that is whether intermediate steps are correct and adhere to standard mathematical methods, irrespective of the final answer. Metrics such as coherence with module content and clarity of response are used to evaluate GenAI responses.

2.2 Sample Size and Selection

2.2.1 Online Maths Questions Bank

The sample of questions comprised those from three online assessments. The first two assessments each contained 10 questions subdivided in several parts. The final and third assessment consisted of 13 questions also subdivided in several parts. An additional dataset file shows the number of questions and subparts in each assessment in more detail [see Additional file 1]. The online test paper is structured with 10 questions, each designed to explore distinct topics covered in the module curriculum. Each question is carefully weighted based on its complexity level. The assessment format includes a variety of question types such as multiple-choice, fill-in-the-blank, and dropdown menus, a few open-ended questions. Students input their responses digitally on the virtual learning environment. The test is conducted online, without invigilation, allowing students to complete it from any location with good and reliable internet access. The duration of the test is set at 1 hour and 30 minutes, during which students have the flexibility to address questions in any order they choose. Automatic submission of the quiz occurs at the end of the allotted time, ensuring equitable assessment conditions and adherence to prescribed examination protocols. In total, 66 questions were tested and redesigned, out of which 24 related to Test 1, 29 to Test 2 and 13 to Test 3 between 2023 and 2024.

2.2.2 Student Population

A purposive sampling technique was employed as the study included all 211 students who took the tests in the academic year 2022/23 and 217 in the following academic year in 2024. Both cohorts had a good command of English with an average IELTS score of 6.6. Students were foundation level students for one year before joining their preferred undergraduate studies. They were either on the Economic & mathematics or Management & Business pathway at the university.

2.2 Focus Group

2.2.1 Participants

Participants for the focus group were recruited using self-selection sampling as they were invited to participate through a combination of in-person and digital ways. Flyers were posted on the module's MS Teams channel and were shared during lectures and seminars. All students enrolled on the module were aware of the opportunity to participate. Five students volunteered to attend the focus group, which points to a voluntary and self-selection sampling method, popular in qualitative research (Bryman 2008). Each student was sent a Participant Information Sheet which contained all necessary information about the study, including what happens if they take part, what if they change their minds, data handling, data confidentiality and any risks. Participants were then asked to fill a consent form. The focus group was conducted online via MS Teams at the end of the academic year and lasted for one hour. The online setting was chosen for its accessibility, especially that students already left the country. Moreover, rigour of the focus group was maintained which ensured that responses could be accurately transcribed and reviewed.

Semi-structured questions, that included open-ended questions to encourage broad discussion and more probing questions to delve into specific areas like clarity and difficulty were asked and related to students' overall experience with online assessments. Several areas were explored: students' perceptions of the redesigned GenAI-resistant questions, approaches students used to answer them as accurately as possible in an online setting, challenges they faced during the tests, the use of GenAI tools following the tests and suggestions for improvement when taking tests/quizzes on an online platform.

To foster an open, non-judgemental, comfortable environment and ensure data reliability, students had the option to participate either in writing by using the chat function or verbally through video, with the freedom to switch their cameras on or off. Participation in discussion was voluntary, and students were free to leave at any time. Students were assured that there was no 'wrong' or 'right' answers and questions were phrased in a neutral tone to encourage honest responses, especially with the use of GenAI. Some follow-up questions were spontaneous, emerging from students' responses. That encouraged a natural flow of conversation and contributed to the depth and data richness (Bryman 2008).

The focus group was also a good platform for students to reflect on and share their honest use (or non-use) of GenAI tools after the assessments, which shed light on how students were engaging with the new technology and how it impacted their learning.

2.2.2 Reflexive Thematic Analysis

Transcripts[2] were anonymised without distorting scholarly meaning and analysed using the six-phase TA approach detailed by (Braun and Clarke 2006). Through multiple phases, main themes and sub-themes were systematically identified and interpreted. NVivo 14 (Lumivero 2023) package was used for data management and coding. Data were further independently coded by a second researcher to maintain the reliability and validity of themes identified and reduce researcher bias (Silverman 2016).

3. Findings and Data Analysis

3.1 ChatGPT-Proof Question Types Analysis

Table 1 summarizes the total number of questions answered incorrectly by ChatGPT (referred as ChatGPT-proof") for each of the online tests, pre and post redesigning the questions to make them more GenAI-resistant.

Table 1: ChatGPT-proof questions: Pre and Post Test Comparison

Online Test	**ChatGPT Proof**		
	Pre	**Post**	**Total**
1	24	31	55
2	29	34	63
3	13	24	37
Total	66	89	155

Overall, the number of questions considered resistant to ChatGPT increased across all three tests, rising by 23 in total. With respect to the strategies implemented to enhance the effectiveness of questions against GenAI, table 2 reports the changes across the three online test questions pre and post ChatGPT which are either ChatGPT-proof or not.

[2] Available upon request.

Table 2: Impact of ChatGPT-proof Strategies on Online Test Questions

Online Test	**ChatGPT Proof**			
	No		**Yes**	
	Pre	**Post**	**Pre**	**Post**
1	17 71%	5 16%	7 29%	26 84%
2	23 79%	8 24%	6 21%	26 76%
3	8 62%	7 29%	5 38%	17 71%

The strategies worked and enhanced the robustness of questions against ChatGPT. Test 1 questions increased from 29% to 84%, test 2 experienced a notable rise of 55% in ChatGPT-proof questions while test 3 showed a moderate improvement from 38% to 71%, likely because some questions were already ChatGPT-proof pre-redesign.

To further understand these improvements, Table 3 provides a detailed breakdown of the proportion of ChatGPT-proof questions by question type, offering deeper insights into how they contributed to the overall effectiveness of the strategies. As highlighted in the literature, higher-order thinking questions are less susceptible to being effectively answered by GenAI tools like ChatGPT (Farrelly and Baker 2023). This is clearly reflected in Table 3, which demonstrates the effectiveness of different question types in becoming ChatGPT-proof across online tests.

Table 3: Question Type Performance Against ChatGPT Across Online Tests

		Pre		Post	
Online Test	**Question Type**	**No. of Questions**	**% ChatGPT Proof**	**No. of Questions**	**% ChatGPT Proof**
1	Recall	5	0%	1	0%
	Understanding	2	0%	2	50%
	Application	7	14%	8	75%
	Analysis	7	43%	4	100%
	Evaluation	0	0%	6	83%
	Creation	3	100%	4	100%
	Graph	0	0%	6	100%
2	Recall	6	0%	3	67%
	Understanding	0	0%	1	0%
	Application	13	8%	8	38%
	Analysis	7	43%	5	80%
	Evaluation	3	67%	8	100%
	Creation	0	0%	4	100%
	Graph	0	0%	5	100%
3	Recall	1	0%	0	0%
	Understanding	0	0%	1	0%
	Application	8	25%	10	40%
	Analysis	1	0%	0	0%
	Evaluation	2	100%	7	100%
	Creation	0	0%	3	100%
	Graph	1	100%	3	100%

In Test 1, questions that require higher-order thinking, such as "Evaluation" and "Creation," consistently achieved 100% ChatGPT-proofness. Specifically, "Application" questions saw a significant improvement, rising from 14% to 75%, while "Analysis" questions increased from 43% to 100%. These results align with the literature's assertion that such questions are less likely to be handled well by GenAI. In Test 2, the most notable changes were observed in “Recall”, “Application,” and again “Evaluation” questions. ChatGPT-proof recall questions increased from 0% to 67%, and application questions moved from 8% to 38%. The improvement in these question types indicates a successful redesign in making even basic recall questions more resistant to GenAI-generated responses (Weimer 2018). For Test 3, "Application" questions saw an increase from 25% to 40% ChatGPT-proof questions. Newly introduced "Creation" questions were 100% ChatGPT-proof as expected. This consistency highlights the effectiveness of the redesign in maintaining high standards for question robustness across different test formats.

3.2 Students' Perceptions of Online Tests

A total of 6 main themes were generated based on the reflexive thematic analysis to capture the recurrent thoughts common across the students.

Theme 1: Question Types

A variety of question types were designed to assess different skills, such as recall, application, and critical thinking. More than half of respondents preferred multiple choice and fill-in-the-blank questions, which were perceived as easier. Open-ended questions were more challenging (Reardon et al. 2018):

> P4: I mean multiple choice was my top choice. Obviously, it's easier than the others and the fill-in-the-blanks was also easy too. […] There was some open-ended questions, right? Alright, so I had a hard time solving them. I think 'cause [because] there were too many steps.

However, one participant (P1) suggested including more real-world problems, which they described as 'more creative'. These higher-order thinking questions were appreciated for their connection between maths and social sciences. Another participant (P5) expected more challenging questions, given the 'open book' test setting.

> P5: I knew that all of the tests are going to have that sort of question, that sort of flavour. So I was expecting this thing and yeah, I think these are the questions […] for an online test or Open book test, it's usually would become like this.

Theme 2: Clarity and Difficulty

Majority of the students found 'the questions in the test three was really challenging' (P1) compared to Tests 1 and 2. They reported that the questions were more confusing and 'not connected with materials covered in class' (P3), which made them more time-consuming to answer.

> P2 [with respect to test 1]: "I think the test one while being a little bit confusing, there was still kind of straightforward questions. We could easily identify why the question was asking us and what we needed to do to solve them."

This could be due to the complexity of the questions, the way they were phrased, or the topics they covered to make them more GenAI-resistant. Students understood 'it was trying to be like AI softwares and everything' (P2), but believed 'they did become more confusing in test 3.'

Overall, everyone except one (P3) that the three tests varied significantly in difficulty, with the test 1 being intermediate, test 2 easy, and test 3 unexpectedly hard.

P4: The three tests were all unexpected because the first Test was. I feel like it was intermediate. […] the 2nd test was really easy. It was just really easy. And the third one was really hard, so there was a big difference between the three."

Theme 3: Preparation and Strategies or Feedback and Adaptation

All participants used various strategies to prepare for the tests, such as reviewing lecture slides, watching videos, and practicing exam-style questions. However, some students felt that these strategies were not as effective for Test 3 due to its increased difficulty and lack of clarity. P4 mentioned that 'my study routine was the same for all three tests, but the third one was my lowest grade'.

However, 2 out of 5 students adapted their study strategies over time, but some felt that better preparation might not have significantly improved their performance 'because it [test 3] was really hard for me' (P4).

The design of the tests allowed for feedback and adaptation (Brown 2022). After each test, students could review their performance and adjust their study strategies for the next test.

P5: I made some mistake in the first Test that I tried to maybe study the textbook or cover some things that wasn't emphasised in the workbook and Desmos; so on the 2nd test I tried to just focus on the main topics that was in the workbook and the main topics. All the questions that were given to us in Desmos the questions that we had to complete before the seminar lecture and that sort of things. I think I did well. For the third test I adapted the same approach.

Theme 4: Use of GenAI post-tests submission

There was a consensus that the use of GenAI, especially ChatGPT was not effective in correctly solve math questions and provided some formulas which were not relevant to the tests' questions.

P5: […] I think generally ChatGPT is not good with maths like it always gives you weird questions. So I knew that even if I tried it would give me a wrong answer so I didn't try honestly.

Another student indicated that they did not use ChatGPT but would use feedback provided 'below each question [there's like a yellow box]' (P2), suggesting a preference for the solutions provided by the module team rather than depend on external GenAI tools.

Two respondents clearly stated that they 'put the questions' (P5) into ChatGPT after the tests were completed but the unsatisfactory results led them to avoid using it for the other tests.

However, P4 believed that their lack of access to the 'fourth version' might offer better functionality, although the basic versions used were deemed inadequate.

Theme 5: Technical Issues

All participants took the tests in their own accommodation, making sure that there is a ‘constant internet connection’ (P1, P2). Additionally, no respondent faced technical difficulties during the time allocated for filling in their answers online.

P5: I didn’t have any glitches, technical issues during the test time […] and not any bugs or anything like that. It, was uh, it was smooth

However, P2 mentioned ‘a lot of friends’ experiencing issues with the online platform, particularly when it came to submitting their workings (after the quiz has been answered, students were given an extra 30 minutes to upload their notes).

Theme 6: Preference for On-Campus Tests

It was unanimous among students that the university should ‘set more on campus tests so students can take exams [tests] fairly (P3) (Henderson et al. 2022). They believed that setting on campus tests will prepare them for the final on campus exams, 3 of whom referred to the mental preparation.

P1: They’re subconsciously preparing for the mentality they have to have a mock exam so they don’t get more nervous.

P5: King’s needs to change to have the test on campus as soon as possible […] It can prepare how to control their [students] nerves when the tests are on campus.

Most participants raised the issue with cheating in an online setting where ‘some students are gathering like in groups and solving it [the questions] together (P4). One student explained how taking the online test alone in their room lead to more mistakes.

P3: […] because in online [tests], just myself in my room so I always make small mistakes like calculate [calculations] but I also take mock exam on campus […] I calculate more accurately, so maybe feel some anxiety.

Finally, taking tests on campus also meant test questions ‘being more direct’ and ‘easier to understand’ (P2).

3.3 Student Grade Performance Trends and Comparative Analyses Across Cohorts

This section examines student performance over several years, beginning with the introduction of online tests in January 2020 during the COVID-19 pandemic. Significant changes were implemented in the 2023-24 academic year to address the increasing prevalence of GenAI tools like ChatGPT, which began gaining traction around January 2023. While online tests were already in use, their design evolved over time. In the 2020-21 academic year, the test questions remained largely unchanged. The 2021-22 year saw some adjustments, though these were not substantial. High scores in the initial years, combined with challenges identified in online testing, led to a redesign of questions for the 2022-23 year. This included generating multiple versions of questions, increasing the use of word problems, and reshuffling question formats. However, by 2023-24, it became clear that many questions, particularly those focused on recall, understanding, and application, were too easily answered by ChatGPT. Consequently, a major redesign was undertaken to enhance question difficulty and reduce GenAI predictability.

Table 4 highlights the impact of the 2023 redesign on student performance. A downward trend in scores was evident even before the redesign, following some question adjustments and strategy changes in the 2022/23 academic year.

Table 4: Trends in Students' Performance in Online Tests and Normality Test (2020-2024)

Online Test 1	**2020-21**	**2021-22**	**2022-23**	**2023-24**
Mean	86.4	80.3	65.8	69.5
Standard deviation	13.8	14.5	16.7	18.2
Median	90.0	83.0	66.0	71.0
Shapiro-Wilk statistics	0.808***	0.898***	0.968***	0.841***
Sample size	166	203	211	227

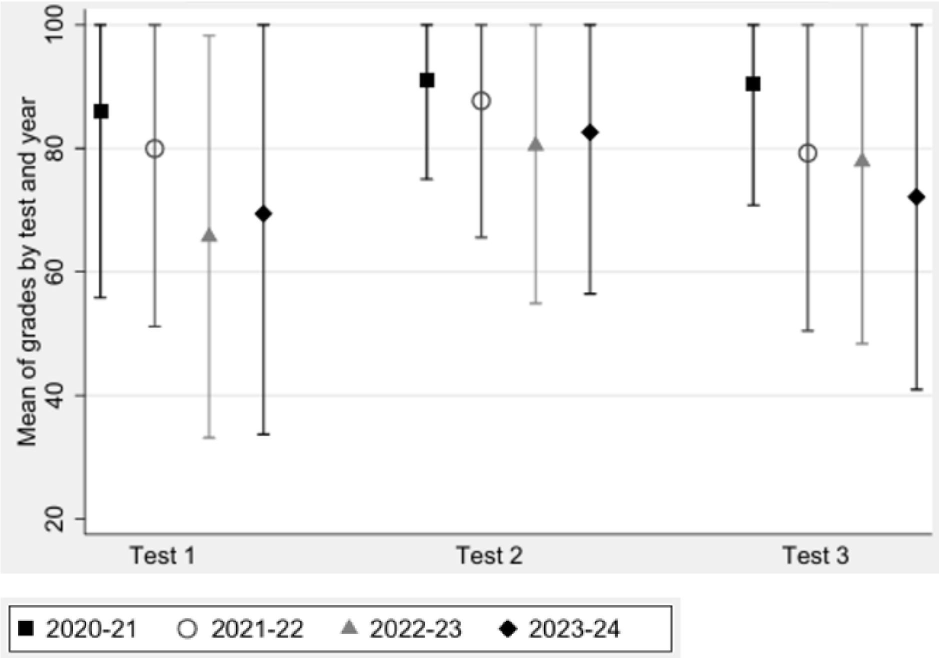
Online Test 2	**2020-21**	**2021-22**	**2022-23**	**2023-24**
Mean	91.0	88.4	80.6	82.8
Standard deviation	8.2	10.6	13.0	13.3
Median	93.0	90.0	83.0	85.0
Shapiro-Wilk statistics	0.841***	0.781***	0.953***	0.914***
Sample size	166	203	211	227

Online Test 3	**2020-21**	**2021-22**	**2022-23**	**2023-24**
Mean	90.6	79.8	77.8	72.2
Standard deviation	10.0	13.7	15.1	15.9
Median	93.5	83.0	82.0	73.0
Shapiro-Wilk statistics	0.807***	0.891***	0.949***	0.975***
Sample size	166	203	211	227

*** significant at 1% level.

Grades for Test 1 and Test 2 declined monotonically up to 2023 when there was a slight reversal post ChatGPT. However, Test 3 grades experienced a further decline in 2024, which question its level of difficulty. Standard deviations in Table 4 also reflected that test grades varied across cohorts. To account for this variability or precision of the mean estimate, confidence intervals are reported in **Fig.** 9.

Fig. 9 confidence intervals of student grades 2020-2024



Test 1 and 3 grades displayed wider confidence intervals indicating more uncertainty in the mean estimates caused by higher variation in performance. Narrower confidence intervals for test 2 in recent years suggest that mean grades stabilised or fluctuated less. Most confidence intervals appeared to be overlapping, implying no significant differences between time periods except possibly for tests 1 and 2 in 2022/23. However, confidence intervals may be unreliable if the underlying grades do not follow a normal distribution. Shapiro-Wilk tests confirm the rejection of the null of normality at 1% level of significance in all cases, as per Table 4. Therefore, to ensure the robust comparisons of differences in means across cohorts and over time, the Welch's t-test is used, which also take into account unequal variances. The findings pre and post ChatGPT periods across the three assessments are reported in Table 5.

Table 5: Welch's t-tests of student grades pre-ChatGPT and post-ChatGPT

Online Test 1	**Comparison Years**	**t-statistics**	**p-value**
Pre-ChatGPT	2021-2022	4.135***	0.000
	2022-2023	9.454***	0.000
Post-ChatGPT	2023-2024	-2.244**	0.025

Online Test 2	**Comparison Years**	**T-statistics**	**p-value**
Pre-ChatGPT	2021-2022	2.657***	0.008
	2022-2023	6.734***	0.000
Post-ChatGPT	2023-2024	-1.722*	0.086

Online Test 3	**Comparison Years**	**T-statistics**	**p-value**
Pre-ChatGPT	2021-2022	8.678***	0.000
	2022-2023	1.411	0.159
Post-ChatGPT	2023-2024	3.830***	0.000

***, **, * significant at 1%, 5%, 10% level respectively.

There were significant improvements in grades for test 1 between 2021-22 (t=4.135) and 2022-23 (t=9.454), representing increases in academic performance in the pre-ChatGPT period. However, due to more GenAI-resistant questions, there was a drop in academic performance in the post-ChatGPT period with t=-2.244, significant at 5% significance level. Similar trends were observed for test 2 but post ChatGPT saw a marginal decline in overall performance, significant at 10% level. The most differences were observed for test 3 with highly significant differences in 2021-2022 period (t=8.678) while 2022-2023 showed no significant difference. Post ChatGPT, there was again a highly significance decrease in academic performance.

4. Discussion and Limitations

4.1 Online Question Design

The question type analysis highlights that "Understanding" and "Recall" questions were less effective as ChatGPT would easily give the correct answer. These question types either saw minimal or no change in their proofness rates post-redesign, with "Understanding" and "Recall" questions remaining largely unchanged or being reduced in number after the redesign process. Overall, the redesign efforts notably enhanced the proofness of "Application" and "Analysis" questions, demonstrating their effectiveness in challenging GenAI tools (Salinas-Navarro et al. 2024). The consistently high proofness of "Evaluation" and "Creation" questions further underscores their robustness against GenAI, while "Understanding" and "Recall" questions proved less effective in this regard. These findings underscore the importance of incorporating higher-order thinking questions in assessments to mitigate the impact of GenAI tools like ChatGPT and demonstrates the relevance of Bloom's higher order thinking skills approach to assessment design. The significant improvements in certain question types validate the strategic approach to redesigning online tests, ultimately leading to more robust and rigorous assessments.

ChatGPT excels at solving straightforward problems. Educators can create assessments that test deeper understanding and critical thinking rather than rote calculation using the following key strategies. They have been found to challenge GenAI in the context of applied maths in the Economics, Finance and Business topics.

Focus on Formula Selection rather than Formula Application

GenAI tools can easily use formulas to solve questions such as interquartile range or demand and supply questions. Instead of asking students to calculate the interquartile range using a formula, the question could rather ask students to figure out which formula would lead to a given outcome. GenAI find this reverse-engineering approach more difficult as it requires deeper engagement and reasoning skills.

Include Conceptual Constraints in Real World Problems

ChatGPT easily handle straightforward mathematical problems, such as finding the intercepts of a quadratic function unless real-world constraints are added. For example, economic outputs cannot be negative or the model struggles when taxes and subsidies are accounted for. Asking students to analyse whether solutions are realistic requires them to think critically, applying economic reasoning rather than just mathematical skills.

Introduce Ambiguity in Theoretical Questions

Questions involving unambiguous retrieval information, like selecting supply and demand equations among a list of options, are to be avoided. Instead, adding complexity through ambiguity or introducing additional, potentially irrelevant information can make these questions more challenging. For the demand equation example, including multiple possibilities and irrelevant data requires filtering and critical thinking.

Review display of Tables and Representation of Data

Rather than using traditional methods of representing data in tables which GenAI can easily handle and interpret, presenting data in a non-traditional way like reversing the tables (if possible) challenges it. An example question is asking students to fill in missing values prompting them to work backward, requiring higher-order cognitive skills.

Work Backwards from the End Solution

Traditionally, students worked through a problem step-by-step to reach a solution, which GenAI can do easily. Instead, by reversing the process and starting with the solution, students can be asked to deduce the steps which engages them more with the material and enhances their problem-solving skills. Such approach is effective in the case of figuring out operations or transformations within systems of equations to arrive at a final equation.

Include Evaluative Statements in True/False Questions

True or False tasks can include more abstract or argumentative statements as compared to simple factual recall ones. For instance, students can be asked to assess whether statements are always true or only under certain conditions which GenAI finds more challenging.

Use Diagrams as a Last Resort

Topics like probability and differentiation can include diagrams or non-standard notation to introduce some level of complexity. For instance, displaying a tree diagram with missing values to answer a probability question can test GenAI's limits in interpreting visual data.

4.2 Focus Group

These perceptions highlight the importance of clear, well-designed questions and a user-friendly test platform in online testing. It also underscores the need for institutions to consider student feedback when designing and implementing assessments.

Surprisingly, most of the participants preferred moving to on campus tests and exams despite the convenience of online testing from educators. This could suggest that the format of the online tests was not as effective or comfortable for them as traditional, on-campus tests, especially with this issue of cheating. They felt that on-campus tests resulted in less stress, had fewer technical issues, and better prepared them for final exams. This feedback suggests that while online testing can offer flexibility, it's important to ensure that it doesn't compromise the quality of the testing experience.

With respect to GenAI usage, although students could access GenAI tools like ChatGPT, they did not trust them enough to use them during the online tests because of the perceived inaccuracy in solving mathematical questions. These perceptions, enhanced by the fact that questions were designed to be more AI-resistant, may indicate an improved understanding of GenAI limitations among students.

Considering student feedback when designing and implementing assessment are important to ensure that the tests are clear, manageable, and free of technical issues to provide a positive testing experience for students. It's also important to provide adequate resources and support to help students prepare effectively for these tests. Issues with submitting workings post quiz submission could indicate problems with the design of the test interface or the functionality of the submission process, suggesting room for improvement in this area.

The design of the online tests was thoughtful and aimed at providing a comprehensive assessment of student learning. However, the unexpected difficulty of the third test indicates areas for improvement. It's crucial to ensure that the tests are not only challenging but also fair, and that the technical aspects of the test administration are smooth and user-friendly.

4.3 Differences in Grades Across Cohorts

Overall, students performed exceptionally well in test 2, with consistently highest mean scores, with tests 1 and 3's mean scores showing greater variability. The differences in confidence interval width across tests confirmed these varying levels of uncertainty. This suggests that students' performance in Test 2 became more reliable, whereas Tests 1 and 3 exhibited more fluctuation. Interestingly, Test 1 saw a slight improvement, corroborated by one of the main themes from the focus group discussion with students. They welcomed the additional support provided during seminars to help students adapt to the new question types (Stamov Roßnagel et al. 2021). Test 2's scores were expected to be higher, as students became more familiar with the question formats and could prepare more effectively. However, this minor drop could suggest that the strategies may not have been as successful, highlighting the difficulty in constantly counteract GenAI capabilities. Test 3's performance drop indicates that students found the revised questions more difficult, and test questions not sufficiently mirroring seminar content. These results highlight that students tend to struggle more on higher-order thinking. It is therefore important to ensure that future question designs are adaptive and continue to effectively balance alignment with module content and challenge.

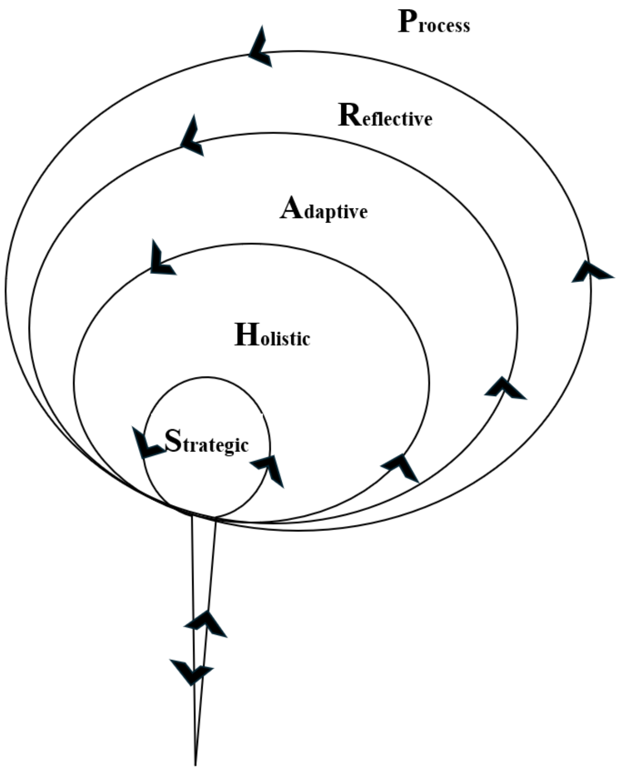
The strategies used in redesigning GenAI-resistant questions appears to be effective, supported by the significant negative effect on academic performance across tests and between 2023 and 2024. This is further reaffirmed by the fact that they did not plummet drastically but reflect a healthy and controlled shift. The new assessment design strategies, although successful, highlights the potential challenges introduced by GenAI in maintaining rigor inclusivity and academic integrity while achieving learning outcomes while creating any sort of assessment.

4.3 The Newly Developed SHARP Assessment Cycle: the 4Rs Approach redefined

The 4Rs approach (Reflect, Rewrite, Retest, Review) started with in this study provided a structured, iterative framework for developing GenAI-resistant assessments in education. It adapted to advancements in GenAI by continuously refining questions, ensuring that assessments stay relevant and effective. The process helped in reflecting on the weaknesses of questions, rewriting them to a higher order thinking level, retesting them against ChatGPT, and reviewing responses to match learning outcomes (Stamov Roßnagel et al. 2021). Overall, this data-driven approach helped the module's instructors not only to fine-tune online tests to make them more GenAI-resistant but also foster deep learning and independent problem-solving skills. Importantly, the review phase ensured fair and pedagogically sound redesigned questions. However, the four stages lack a more student-centred focus that will ensure questions are also accessible. The retesting phase also misses the rigorous checks regarding other GenAI tools to assess their capabilities.

The SHARP assessment cycle, see **Fig.** 10, therefore enhances the existing Reflect-Rewrite-Retest-Review process by addressing the above weaknesses and providing a more holistic framework for developing fair and inclusive assessments in the age of GenAI.

Fig. 10 the SHARP Assessment Cycle



31

SHARP stands for

Strategic: incorporates the careful design of assessments considering GenAI-generated responses. This aspect overviews the rapid progression of GenAI technology by focusing on both immediate and long-term challenges. By expanding the Retest phase, assessments can become resilient across and/or adapt to multiple GenAI tools.

Holistic: includes a variety of assessments (online tests/exams, coursework, presentations, projects) and integrates innovation, student voice and inclusivity (Bain and Jennifer 2010), providing a good balance among GenAI usage, student learning outcomes and feedback for assessments to remain effective and fair.

Adaptive: ensures assessments stay relevant and challenging without being overwhelming in the continuously evolving GenAI capabilities, student performance, and feedback. Difficulty levels of assessments are evaluated, re-evaluated and adapted to the changing educational landscape through the refining of question design.

Reflective: concentrates on the systematic student-centred feedback and assessment outcomes to refine questions and approaches between the Retest and Review phases. This aspect balances GenAI-resistance questions with their accessibility and clarity, ensuring learning objectives are being aligned (Stamov Roßnagel et al. 2021) while at the same time remain effective for learners.

Process: emphasizes a cyclical and iterative approach of refining assessments through retesting, transparent formative assessments and student feedback. Staff and students work collaboratively in various stages developing teamwork and shared ownership.

4.4 Limitations

This study has several potential limitations. One key limitation is the rapidly evolving GenAI landscape during the writing stage, which followed the completion of data collection in July 2024. At the time of writing with a focus on ChatGPT-3.5, advanced iterations of ChatGPT and other artificial intelligence tools, like Gemini, Copilot, Claude, became available, potentially performing better. The effectiveness of the assessment strategies may differ if tested against other AI tools, which may offer better ways to handle complex questions (Hersh and Fultz Hollis 2024). However, according to Digital Education Council Global AI Student Survey in 2024, ChatGPT was the most common used tools by students (66% of 3839 responses across 16 countries). Furthermore, the study presents the 4Rs strategies, a systematic approach to designing more rigorous assessments for economics-related quantitative subjects in the GenAI age. This approach inherently requires ongoing testing of questions using different GenAI tools. Given rapidly changing educational technologies, the SHARP assessment cycle is particularly relevant as it emphasizes an iterative rather than a static process to ensure adaptability.

A restrictive assumption made in this study was that, at the time the tests were carried out, most students had limited access to or under of ChatGPT. Although most students could freely access ChatGPT-3.5, only a few had access to ChatGPT-4 paid version. They may then had access to more sophisticated features which could have enhance their performance on the online assessments.

Another limitation is that the study focuses on first year foundational students in the economics and management areas, hence not fully capturing the diverse skills and knowledge which more advanced students may possess, limiting the generalisability of the findings to other educational settings and disciplines.

5. Conclusion

This paper critically analysed how online assessments are vulnerable in the age of GenAI tools, primarily using ChatGPT-3.5, and reflected on how to transform conventional assessment designs to maintain academic integrity. The Reflect-Rewrite-Retest-Review process provided both a solid framework in redesigning the questions and a continuous reflection on their efficacy and performance. The Bloom's taxonomy was further proven to be effective in significantly reducing GenAI-generated responses to questions which require higher-order thinking skills. Strategies implemented in redesigning the online Maths for Social Sciences tests led to a three-fold increase in ChatGPT-resistant questions across two out of three tests: from 29% to 84% in Test 1, from 21% to 76% in Test 2 and from 38% to 71% in Test 3 (two-fold increase). These findings highlight the importance of shifting from purely recall-based questions to more analytical, evaluative and creative tasks.

However, this transformative approach to assessment design introduced some challenges, especially in terms of student performance in Test 3. There was strongly significant drop in academic performance over the years, suggesting the successful implementation of GenAI-resistant questions but hinting on a rise in the level of difficulty. Student feedback from a focus group confirmed these results whereby students struggled with Test 3's redesigned questions which they said were more complex and lacked clarity as compared to the previous tests. These findings highlight the importance of finding the right balance between designing assessments that meet students' learning outcomes and that challenge GenAI tools while at the same time ensuring that students are sufficiently prepared for tasks that require higher-order thinking levels.

The need to focus on student experience and accessibility and the strategic planning to address advancements in multi-AI models led to the SHARP assessment cycle (Strategic, Holistic, Adaptive, Reflective Process). This comprehensive approach to assessment design is iterative, adaptive, incorporates real-time student and faculty feedback and fosters an inclusive and collaborative learning environment. It offers a practical framework to maintain the balance between with more GenAI-assisted learning experiences while ensuring academic integrity.

In conclusion, this study provides some implications for further research, policy and practice in the context of assessment design in the age of GenAI. Research may incorporate more empirical evidence on how to integrate GenAI tools into assessment designs while balancing knowledge accumulation with the development of higher-order thinking skills to enhance students' positive learning experiences. Moreover, there is need for more empirical studies which test the iterative nature of assessments like the SHARP assessment cycle across diverse assessments and disciplines, adapting them to evolving students' needs and GenAI-enhanced learning environments. Policy implications may involve transforming assessment strategies to be adaptive and dynamic, as new technologies emerge. It is proposed that assessments become more iterative, where they are continuously updated to include students' feedback and GenAI performances, hence upholding academic rigour. Students should be guided to understand GenAI capabilities, limitations and impact on their academic journey. Finally, this paper incorporates some practical steps for educators to co-create assessments with students, fostering a sense of ownership over their learning experience. This collaboration may also raise their awareness about using GenAI in a more responsibly manner and the importance academic integrity, within and beyond the classroom.

List of abbreviations

4Rs: The Reflect-Rewrite-Retest-Review approach

GenAI: Generative Artificial Intelligence

P1: Participant One

P2: Participant Two

P3: Participant Three

P4: Participant Four

P5: Participant Five

SHARP: Strategic, Holistic, Adaptive, Reflective Process

Declarations

Availability of data and materials

The data on question types that support the findings of this study are openly available as supplementary material to the submitted manuscript.

Focus group and performance data for this manuscript are not publicly available but are available upon request.

Competing interests

The author has no competing interests to declare that are relevant to the content of this article.

Author's contribution

Zeenat Soobedar de Villeneuve is the sole author of the study, responsible for its conception and design, methodology, data collection and analysis, the preparation of drafts and the final manuscript.

Acknowledgements

The author would like to thank the Programmes team, especially Adam Best, for their assistance in anonymising the datasets and the students who volunteered for the focus group in King's Foundation. Special thanks to Uzma Razzaque-Bibi for her invaluable guidance in supporting the thematic analysis.

Ethical approval

Ethics approval to conduct this study was obtained from King's College London Research Ethics Committee (LRS/DP-23/24-40829).

Consent to participate

Consent for participation in this study has been obtained according to appropriate ethical standards.

Human Ethics and Consent to Participate

Human Ethics and Consent to Participate declarations: not applicable.

Clinical trial number: not applicable.

Consent for publication

Consent for publication in this study has been obtained according to appropriate ethical standards.

Funding

The author did not receive support from any organization for the submitted work.

References:

Al-Ubaydli O, List JA (2015) On the Generalizability of Experimental Results in Economics. Handb Exp Econ Methodol :420–462. https://doi.org/10.1093/acprof:oso/9780195328325.003.0022

Amzalag M, Shapira N, Dolev N (2021) Two Sides of the Coin: Lack of Academic Integrity in Exams During the Corona Pandemic, Students' and Lecturers' Perceptions. J Acad Ethics 20:243–263

Anderson LW, Krathwohl DR (2001) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman

Bain, Jennifer (2010) Integrating student voice: assessment for empowerment. Pract Res High Educ 4(1):14–29

Bitzenbauer P (2023) ChatGPT in physics education: A pilot study on easy-to-implement activities. Contemp Educ Technol 15(3). https://doi.org/10.30935/cedtech/13176

Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR, others (1956) Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. Longman New York

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3:77–101. https://doi.org/10.1191/1478088706qp063oa

Brown GTL (2022) The past, present and future of educational assessment: A transdisciplinary perspective. Front Educ 7(November):1–8. https://doi.org/10.3389/feduc.2022.1060633

Bryman AE (2008) Of methods and methodology. Qual Res Organ Manag An Int J 3:159–168

Clarke O, Chan WYD, Bukuru S, Logan J, Wong R (2023) Assessing knowledge of and attitudes towards plagiarism and ability to recognize plagiaristic writing among university students in Rwanda. High Educ 85(2):247–263. https://doi.org/10.1007/S10734-022-00830-Y

Clements D, Battista M (1990) Constructivist learning and teaching. Arith Teach 38:34–35. https://doi.org/10.5951/AT.38.1.0034

Digital Education Council (2024). Global AI Student Survey 2024. https://26556596.fs1.hubspotusercontent-eu1.net/hubfs/26556596/Digital%20Education%20Council%20Global%20AI%20Student%20Survey%202024.pdf. Accessed 24 Jan 2025.

Dunne S, Lee D (2022) Designing assessment for neurodiverse students. In: LMJU Student Experience Proceedings (ed) 2022: Students at the Heart Conference. Liverpool John Moores University

Eke DO (2023) ChatGPT and the rise of generative AI: Threat to academic integrity? J Responsible Technol 13:100060. https://doi.org/10.1016/J.JRT.2023.100060

Else H (2023) BY CHATGPT FOOL SCIENTISTS. Nature 613:423

Farrelly T, Baker N (2023) Generative Artificial Intelligence: Implications and Considerations for Higher Education

Practice. Educ. Sci. 13

Flavell JH (1979) Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. Am Psychol 34:906–911

Gikandi JW, Morrow D, Davis NE (2011) Online formative assessment in higher education: A review of the literature. Comput Educ 57(4):2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Henderson M, Awdry R, Chung J, Ashford C, Bryant M, Mundy M, Ryan K (2022) Online exams: exploring student experience and integrity behaviours as we return to campus. ASCILITE Publ :e22095. https://doi.org/10.14742/apubs.2022.95

Hersh W, Fultz Hollis K (2024) Results and implications for generative AI in a large introductory biomedical and health informatics course. npj Digit Med 7(1):1–7. https://doi.org/10.1038/s41746-024-01251-0

Holden OL, Norris ME, Kuhlmeier VA (2021) Academic Integrity in Online Assessment: A Research Review . Front. Educ. 6

Lancaster T, Cotarlan C (2021) Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective. Int J Educ Integr 17(1):1–17. https://doi.org/10.1007/s40979-021-00070-0

Liu D, Bridgeman A (2023) How can I update assessments to deal with ChatGPT and other generative AI? In: 23 January 2023. https://educational-innovation.sydney.edu.au/teaching@sydney/how-can-i-update-assessments-to-deal-with-chatgpt-and-other-generative-ai/

Lumivero (2023) NVivo

Lye CY, Lim L (2024) Generative Artificial Intelligence in Tertiary Education: Assessment Redesign Principles and Considerations. Educ Sci 14(6). https://doi.org/10.3390/educsci14060569

Nikolic S, Daniel S, Haque R, Belkina M, Hassan GM, Grundy S, Lyden S, Neal P, Sandison C (2023) ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. Eur J Eng Educ 48(4):559–614. https://doi.org/10.1080/03043797.2023.2213169

Paul R, Elder L (2013) Critical Thinking: Tools for Taking Charge of Your Learning and Your Life: Pearson New International Edition. Pearson Education

Phillips AJ, Briggs JC, Jensen JL (2019) Beyond Bloom's: Students' Perception of Bloom's Taxonomy and its Convolution with Cognitive Load. J Psychol Res 1(01):24–32. https://doi.org/10.30564/jpr.v1i01.421

Piaget J (1976) To Understand is to Invent: The Future of Education. Penguin Books

Rasul T, Nair S, Kalendra D, Robin M, Santini F de O, Ladeira WJ, Sun M, Day I, Rather RA, Heathcote L (2023)

The role of ChatGPT in higher education: Benefits, challenges, and future research directions. J Appl Learn Teach 6(1):41–56. https://doi.org/10.37074/jalt.2023.6.1.29

Reardon SF, Kalogrides D, Fahle EM, Podolsky A, Zárate RC (2018) The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. Educ Res 47(5):284–294. https://doi.org/10.3102/0013189X18762105

Reedy A, Pfitzner D, Rook L, Ellis L (2021) Responding to the COVID-19 emergency: student and academic staff perceptions of academic integrity in the transition to online exams at three Australian universities. Int J Educ Integr 17(1):9. https://doi.org/10.1007/s40979-021-00075-9

Roe J, Perkins M, Ruelle D (2024) Understanding Student and Academic Staff Perceptions of AI Use in Assessment and Feedback. arXiv Prepr arXiv240615808 :1–13

Salinas-Navarro DE, Vilalta-Perdomo E, Michel-Villarreal R, Montesinos L (2024) Using Generative Artificial Intelligence Tools to Explain and Enhance Experiential Learning for Authentic Assessment. Educ Sci 14(1). https://doi.org/10.3390/educsci14010083

Sallam M, Salim NA, Barakat M, Ala'a B (2023) ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J 3(1)

Seo K, Tang J, Roll I, Fels SS, Yoon D (2021) The impact of artificial intelligence on learner–instructor interaction in online learning. Int J Educ Technol High Educ 18

Silverman D (2016) Qualitative Research. SAGE Publications

St-Onge C, Ouellet K, Lakhal S, Dubé T, Marceau M (2022) COVID-19 as the tipping point for integrating e-assessment in higher education practices. Br J Educ Technol 53(2):349–366. https://doi.org/10.1111/bjet.13169

Stamov Roßnagel C, Lo Baido K, Fitzallen N (2021) Revisiting the relationship between constructive alignment and learning approaches: A perceived alignment perspective. PLoS One 16(8):1–13. https://doi.org/10.1371/journal.pone.0253949

Su J, Yang W (2023) Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education. ECNU Rev Educ 6(3):355–366. https://doi.org/10.1177/20965311231168423

Sweller J, Ayres P, Kalyuga S (2011) Cognitive Load Theory. Springer New York

Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, Kalpathy-Cramer J, Lum F, Kim JE, Baxter SL, Ting DSW (2023) Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. Ophthalmol Sci 3(4):100394. https://doi.org/https://doi.org/10.1016/j.xops.2023.100394

UCL Assessment Working Group (2020) Designing Effective Online Assessment

Van Dis EAM, Bollen J, Zuidema W, Van Rooij R, Bockting CL (2023) ChatGPT: five priorities for research. Nature 614(7947):224–226

Vellanki S, Mond S, Khan Z (2023) Promoting Academic Integrity in Remote/Online Assessment – EFL Teachers' Perspectives. Teach English as a Second or Foreign Lang 26:1–20. https://doi.org/10.55593/ej.26104a7

Vygotsky LS (1980) Mind in Society: The Development of Higher Psychological Processes. Harvard University Press

Weimer M (2018) Multiple-Choice Tests: Revisiting the Pros and Cons. Fac. Focus

Additional files

Test	Topic	PreQuestion	PrePart	PreType	PreChatGPT	PostQuestion	PostPart	PostType	PostChatGPT
1	Linear model	1	a	3	0	1	a	5	1
1	Linear model	1	b	4	0	1	b	5	0
1	Simultaneous Equations 2x2	2	a	6	1	2	a	6	1
1	Simultaneous Equations 2x2	2	b	6	1	2	b	6	1
1	Raw data - measures	3	a	1	0	3	a	6	1
1	Raw data - measures	3	b	6	1	3	bi	3	0
1	Raw data - measures	3	c	1	0	3	bii	3	1
1	Raw data - measures	3	d	1	0				
1	Raw data - measures	3	e	3	1				
1	Equilibrium	4	a	1	0	4	a	5	1
1	Equilibrium	4	b	3	0	4	b	7	1
1	Equilibrium	4	c	4	1	4	ci	5	1
						4	cii	6	1
1	Index numbers and inflation	5	a	3	0	5	a	3	1
1	Index numbers and inflation	5	b	4	1	5	b	4	1
1	Quadratic and graph	6	a	1	0	6	a	7	1
1	Quadratic and graph	6	b	4	0	6	b	7	1
1	Quadratic and graph	6	c	4	0	6	c	7	1
1	Quadratic and graph	6	d	2	0	6	d	7	1
						6	e	7	1
1	Grouped data - measures					7	a	5	1
1	Grouped data - measures					7	b	1	0
1	Grouped data - measures					7	c	3	1
1	Grouped data - measures					7	d	2	1
1	Grouped data - measures								
1	Quadratic and revenue	8	a	4	1	8	a	4	1
1	Quadratic and revenue	8	b	2	0	8	b	2	0

1	Quadratic and revenue	8	c	3	0	8	c	3	1
1	Simultaneous equations 3x3	9	a	3	0	9	a	5	1
1	Simultaneous equations 3x3	9	b	4	0	9	b	4	1
1	Simultaneous equations 3x3	9	c	3	0	9	c	3	1
1	Percentages					10	a	3	0
1	Percentages					10	b	3	1
1	Percentages					10	c	4	1

Test	Topic	PreQuestion	PrePart	PreType	PreChatGPT	PostQuestion	PostPart	PostType	PostChatGPT
2	Paired representative stats	1		1	0				
2	Histogram	2	a	5	1	2	a	1	1
2	Histogram	2	b	1	0	2	b	1	0
2	Histogram	2	c	1	0	2	c	5	1
2	Histogram	2	d	1	0	2	d	5	1
2	Optimising profit	3	a	3	0	6	a	6	1
2	Optimising profit	3	b	3	0	6	b	3	0
2	Probability combined events	4	a	3	0	3	a	3	0
2	Probability combined events	4	b	4	0	3	b	5	1
2	Probability combined events	4	c	4	0	3	d	5	1
2	Probability combined events	4	d	4	0	3	d	3	0
2	Grouped data - measures					8	a	3	1
2	Grouped data - measures					8	bi	2	0
2	Grouped data - measures					8	bii	1	1
2	Grouped data - measures					8	bii	3	1
2	Grouped data - measures					8	biii	3	0
2	Differentiation for graphs	6	a	3	0	5	a	5	1
2	Differentiation for graphs	6	b	3	0	5	b	5	1

2	Differentiation for graphs	6	c	3	0	5	c	5	1
2	Differentiation for graphs	6	d	3	0	5	d	3	1
2	Differentiation for graphs	6	e	3	0	5	e	6	1
2	Differentiation for graphs	6	f	3	0	5	f	6	1
2	Differentiation for graphs	6	g	3	0	5	g	4	1
2	Tree diagram	7	a	3	0	10	a	3	0
2	Tree diagram	7	b	4	1	10	b	4	1
2	Tree diagram	7	c	4	1	10	c	4	1
2	Tree diagram	7	d	4	1	10	d	4	1
2	Tree diagram	7	e	4	0	10	e	4	0
2	Bar chart	8	a	5	0	1	a	7	1
2	Bar chart	8	b	1	0	1	b	7	1
2	Bar chart	8	c	1	0	1	c	7	1
2	Bar chart	8	d	5	1	1	d	7	1
2	Differentiating polynomials	9	a	3	0	9		6	1
2	Differentiating polynomials	9	b	3	1	9			
2	Sketch critique					4		7	1
2	Five-summary statistics table					7		5	1

Test	Topic	PreQuestion	PrePart	PreType	PreChatGPT	PostQuestion	PostPart	PostType	PostChatGPT
3	Modelling with exponential	1	a	3	0	5	a	7	1
3	Modelling with exponential	1	b	1	0	5	b	5	1
3	Separting log terms	2		3	0	2		3	1
3	Normal distribution	3	a	3	0	12	a	5	1
3	Normal distribution	3	b	3	0	3	b	3	1
3	Normal distribution	3	c	5	1	3	c	6	1
3	Index/log form conversion	4		3	0	4		3	1

3	Identify log and exp sketches	5		7	1	5		7	1
3	Normal distribution (mean and standard deviation)	7		4	0				
3	Convert to single log form	8		5	1	1		5	1
3	Product, chain and quotient rule					6	a	3	0
3	Product, chain and quotient rule					6	b	3	0
3	Solving system of equations with matrix					7	a	3	0
3	Solving system of equations with matrix					7	b	5	1
3	Determinant of 3x3 matrix					8		5	1
3	3x3 matrix cofactor					9		5	1
3	Bayes theorem and tree diagram	10	a	3	0	10	a	6	1
3	Bayes theorem and tree diagram	10	b	3	1	10	b	7	1
3	Bayes theorem and tree diagram	10	c	3	1	10	c	5	1
3	Expected value					11	a	3	0
3	Expected value					11	b	3	0
3	Binomial distribution					13	a	2	0
3	Binomial distribution						b	6	1
3	Binomial distribution						c	3	0
3	Binomial distribution						d	3	1

Title of data: Total number of questions and subparts pre and post ChatGPT testing phase

Description of the data: The tables provide details on the types of questions including their subparts for online test 1, 2, and 3. Questions and subparts were grouped according to the revised Bloom’s Taxonomy educational objectives:

Recall, Understanding, Application, Analysis, Evaluation and Creation questions, and an extra type referred to as Graphs.

Test column: test number 1, 2 or 3

Topic column: Topic examined

PreQuestion column: Question number before using the 4Rs approach to make it harder for ChatGPT-3.5 to solve

PreType column: Question type before using the 4Rs approach to make it harder for ChatGPT-3.5 to solve where Question Type is categorised into 7 groups as follows

1: Recall

2: Understanding

3: Application

4: Analysis

5: Evaluation

6: Creation

7: Graphs

PreChatGPT column: Whether ChatGPT-3.5 answered the question correctly before using the 4Rs approach to make it harder for ChatGPT-3.5 to solve (0: ChatGPT answered correctly; 1: ChatGPT answered incorrectly)

PostQuestion column: Question number after using the 4Rs approach to make it harder for ChatGPT-3.5 to solve

PostType column: Question type after using the 4Rs approach to make it harder for ChatGPT-3.5 to solve (see above for Question type categories)

PostChatGPT column: Whether ChatGPT-3.5 answers correctly after using the 4Rs approach to make it harder for ChatGPT-3.5 to solve (0: ChatGPT answered correctly; 1: ChatGPT answered incorrectly)