

Statistics assessment, generative AI and evaluative judgement

Amanda Shaker, La Trobe University

a.shaker@latrobe.edu.au

IMA, LMS & RSS Higher Education Teaching and Learning Workshop
Series 2024/25: *Rethinking Teaching and Learning in The
Mathematical Sciences in Times of Generative AI*

18 July, 2025



- ① Motivation
- ② Pedagogical considerations for assessment
- ③ Teaching and assessment context(s)
- ④ Tools for writing questions
- ⑤ A model for question design using evaluative judgement of GenAI outputs, with 2 examples
- ⑥ Discussion

- Generative Artificial Intelligence (GenAI) has arrived and is here to stay!
- GenAI can be a valuable learning tool for students
- It can also undermine student learning if not used appropriately
- Students will likely be expected to use GenAI once they enter the workforce
- Thus it is important we educate students to use GenAI appropriately and critically evaluate its outputs
- In light of this, there is a need to rethink assessment in Higher Education

Pedagogical considerations for assessment

- Bloom's taxonomy (Bloom, et al., 1956) and MATH Taxonomy (Mathematical Assessment Task Hierarchy) (Smith et al., 1996) can be referenced to consider designing questions that incorporate *Higher order thinking*
- For example, questions that require **evaluation**, **analysis**, **justification**, etc.
- This can be achieved via many question types, including **multiple choice** and **numerical answer** questions.
- For example, download a data set, carry out statistical analysis, and select or state the correct interpretations, justifications or outputs

- **Randomisation** can ensure students receive individualised questions, disrupting copying and collusion in unsupervised assessments
- **Evaluative judgement** is 'the capability to judge the quality of work of self and others' (Bearman et al., 2024, p. 893)
- Questions incorporating *randomisation* and *evaluative judgement of GenAI outputs* achieve the following:
 - ① Incorporate higher order thinking
 - ② Foster the skill of critically evaluating GenAI outputs
 - ③ Disrupt inappropriate use of GenAI in assessment
 - ④ Disrupt collusion and copying in unsupervised assessments

Teaching and assessment context

- Large, introductory first-year subject
- Assessments:
 - 1 Six online quizzes (30%). Open for 1 week each, no time limit, two attempts per quiz, best 5 of 6 quiz marks count
 - 2 Three assignments worth 15% each
 - 3 Final summative assessment; time-based, non-invigilated, worth 25%
- Examples discussed here were developed within this context, and therefore make use of features such as randomisation and automated marking
- This facilitates efficiency, flexibility and scalability while disrupting copying and collusion
- However, many of the concepts and pedagogical considerations can be applied in other contexts, including other assessment types, year-levels, and disciplines

Tools for writing questions

- I used R/exams (Zeileis, 2014) and Moodle quizzes
- R/Exams is an R package, `exams`, allowing dynamic creation of randomised questions
- Allows for either R Markdown or LaTeX syntax, and facilitates randomisation, including randomised data sets that students download and analyse
- Question types include single and multiple choice, numeric and string response, or any combination of these ('cloze'): these can all be computer-marked
- Can also require longer answers or file uploads ('essay' question) for more in-depth responses but requiring manual marking
- R/Exams can be used with other LMS platforms, e.g. Canvas, Blackboard, etc. and to create PDF, HTML or Word documents
- Other e-assessment options include Dewis, Numbas, STACK, etc.

A suggested framework for question design

Part 1. Scenario:

Suppose a GenAI tool was used to help answer the following question:

Part 2. Question to GenAI:

This may be a traditional assessment question or an existing question that is being converted into this new framework.

Part 3. GenAI Answer:

Here, an actual or hypothesised answer from GenAI is provided.

Part 4. Question to you:

This is a question or series of questions that students answer using *evaluative judgement* of the GenAI output provided.

Example 1

Scenario:

The Copilot AI tool was used to help analyse some data for part of an STM1001 Assignment question. In the following question to Copilot, assume Copilot has already been provided with the data in the data file `df.csv`

Data set to be downloaded and analysed

Question to Copilot:

Use the data in the .csv file to answer the following questions:

1. What is the sample mean of cholesterol?
2. What is the sample standard deviation of cholesterol?
3. What is the standard error?

Copilot answer:

To answer your questions:

1. The sample mean of cholesterol is 232.97.
2. The sample standard deviation of cholesterol is 39.29.
3. The standard error is 7.12.

Please note that the standard error is calculated as the sample standard deviation divided by the square root of the sample size. In this case, the sample size is 35, so the standard error is $39.29 / \sqrt{35} = 7.12$

Question to you:

Consider each of the following statements. Which of the following statements is/are true? (Note that to be correct, Copilot's answer should be accurate to at least 3 decimal places.)

Select one or more:

- ☐ a. Copilot used the correct sample size value
- ☐ b. The sample standard deviation provided by Copilot was incorrect
- ☐ c. Copilot used the incorrect sample size value
- ☐ d. The standard error provided by Copilot was incorrect
- ☐ e. Copilot used the incorrect formula for the standard error. One or more value(s) used in the formula was/were also incorrect

Alternatives are shuffled, and randomly chosen from a larger list of alternatives

Example 1

- 'Multiple-choice' question - one or more correct options
- Alternatives randomly drawn from larger pool and shuffled
- Students download a data set (potentially randomised)
- Use software to calculate mean, standard deviation and standard error
- Need to compare their answer with GenAI's answer ***and process***
- Although multiple-choice, this question requires a great deal of higher-order thinking

Scenario:

Suppose a GenAI tool was used to help answer the following question:

Question to GenAI:

Suppose that 77 students were asked their height. The results of the survey are displayed below.

Height (cm)	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
[150,160)	10	10	12.99	12.99
[160,170)				
[170,180)				
[180,190)				
[190,200)				
[200,210)	1	77	1.3	100

The sample size and data provided in the table are randomised

Using the above table, answer the following questions:

[Your answers must be accurate to at least 2 decimal places.]

1. What percentage of students' heights are at least 160cm and less than 190cm?
2. What percentage of students' heights are less than 180cm?
3. What percentage of students' heights are at least 170cm?

GenAI answer:

"Using the above table, we can answer the following questions:

1. The percentage of students' heights that are at least 160cm and less than 190cm is the sum of the relative frequencies for the [160,170) and [170,180) height intervals, which is $27.27\% + 41.56\% = \mathbf{69.83\%}$.
2. The percentage of students' heights that are less than 180cm is the cumulative relative frequency for the [170,180) height interval, which is $\mathbf{81.82\%}$.
3. The percentage of students' heights that are at least 170cm is the sum of the relative frequencies for the [170,180), [180,190), [190,200), and [200,210) height intervals, which is $41.56\% + 12.99\% + 3.9\% + 1.3\% = \mathbf{42.87\%}$.

Some randomisation incorporated here

If you need further assistance or have more questions, feel free to ask!"

Question to you:

Consider each of the following statements. Which of the following statements is/are true? (10 marks)

Select one or more:

- ☐ a. For Question 1, the explanation provided by GenAI was incorrect and the final answer provided was also incorrect
- ☐ b. For Question 2, the explanation provided by GenAI was incorrect and the final answer provided was also incorrect
- ☐ c. In total, GenAI provided correct final answers for 2 out of 3 questions
- ☐ d. For Question 1, the explanation provided by GenAI was incorrect but the final answer provided was correct
- ☐ e. The final answer provided by GenAI for Question 3 was incorrect

Alternatives are shuffled, and randomly chosen from a larger list of alternatives

Example 2

- 'Multiple-choice' question - one or more correct options
- Alternatives randomly drawn from larger pool and shuffled
- Output provided with GenAI's interpretations
- The output and where / how the errors occur in GenAI output can be randomised
- Students need to evaluate GenAI's answers ***and process***
- Although multiple-choice, this question requires a great deal of higher-order thinking

- These were just two examples of multiple-choice quiz questions, but the framework can be applied to many other assessment types, question types, and contexts
- This includes in-semester and end-of-semester assessments, with or without invigilation
- In an un-invigilated context, these randomised questions disrupt copying and collusion
- GenAI may be used to assist, but by the nature of the question, doubt is cast over the accuracy of GenAI outputs. In the end, students must still make a decision based on their own judgement
- The questions provided may or may not include mistakes. While GenAI may make one-off mistakes, it is a constantly evolving tool, so the underlying skill of judging its outputs is more important than the specific type of 'mistake' (or lack thereof) in a particular question

- While the skill of evaluative judgement is implicitly taught through this framework, explicit teaching of this skill is also recommended
- This may be seen as a diversion from core content mastery, but: evaluative judgement is an increasingly important skill in a GenAI world; evaluative judgement also requires sufficient understanding of core content, and is therefore one way for this to be demonstrated.
- **Further work:** Evaluating this or other frameworks in various contexts, including different question types, assessment types, year levels and disciplines. Contact me if you would like to collaborate!

References

- Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893-905. <https://doi.org/10.1080/02602938.2024.2335321>
- Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain (1103-1133)*. New York: Longman.
- Smith G., Wood L., Coupland M., Stephenson B., Crawford K., & Ball G. (1996). Constructing mathematical exams to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, 27(1), 65–77. <https://doi.org/10.1080/0020739960270109>
- Zeileis, A., Umlauf, N., & Leisch, F. (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. *Journal of Statistical Software*, 58(1), 1-36. <https://doi.org/10.18637/jss.v058.i01>

Questions

