Quiz Question Generation Using Co-pilot Studio

Embedding Best Practices and Criticality into Design

CAIT Forum 2025

Stewart Utley - Learning Designer
Irina Mylona - Senior Learning Designer
Cambridge Online Education, CUPA

The use-case

Many of our online courses feature **quiz-based comprehension checks**

Stakeholders

Subject Matter Experts

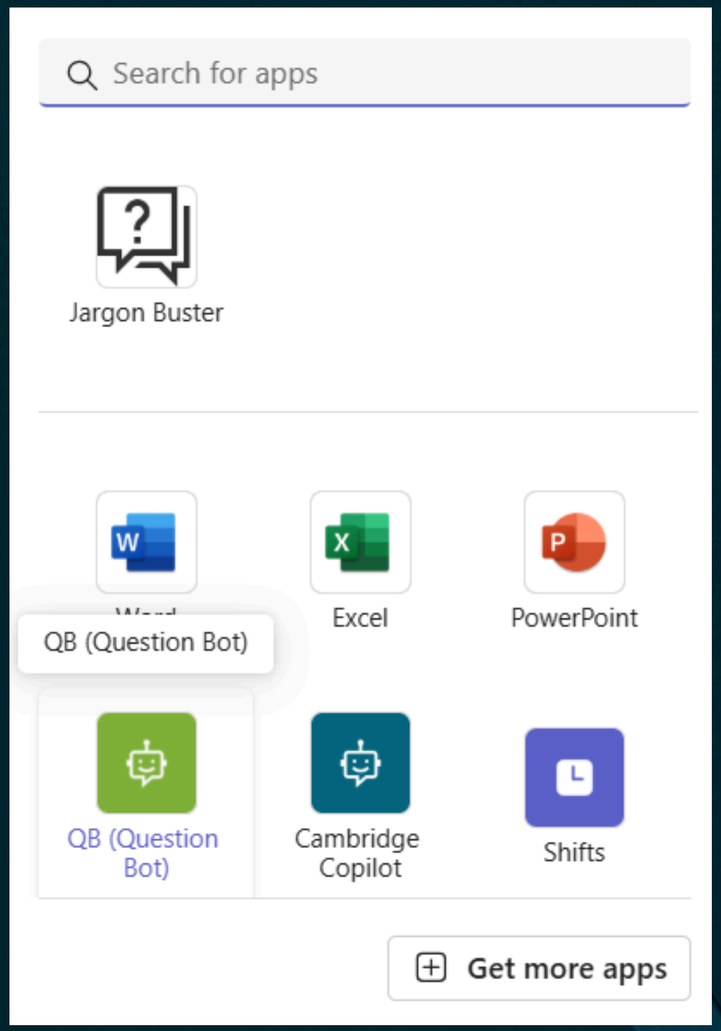
Learning Designers

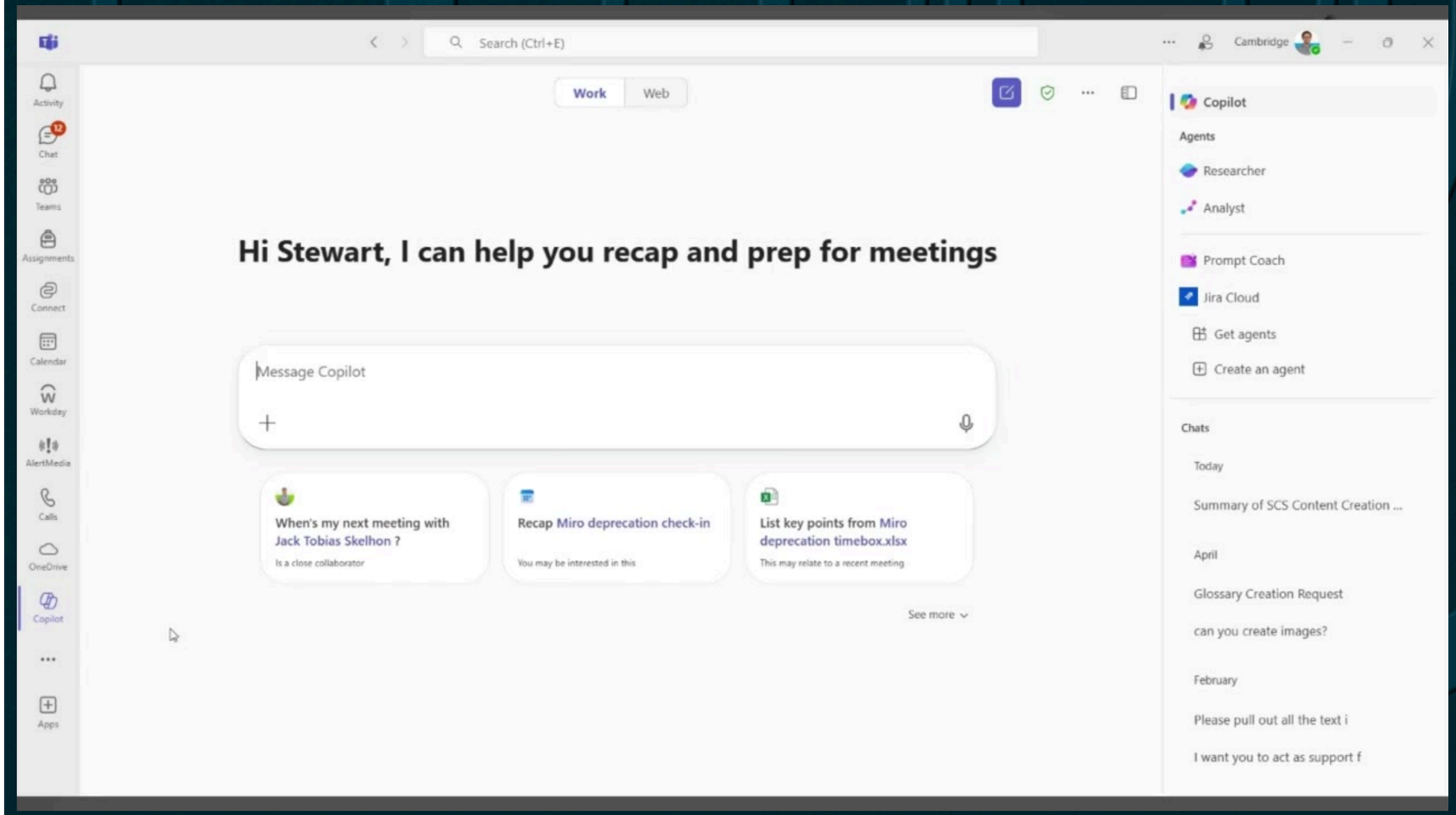
Content Editors

Students

Guiding Question: How can we best utilise Generative AI to optimise the process for creating formative quiz questions whilst maintaining quality?

Access to Question Bot set via. Teams





Prompt engineering

Assessment best practices

Prompting best practices

Role-playing

You are an expert assessment writer. Create ten multiple-choice questions (stem + 4 options). Follow the rules below:

Task clarity

(Haladyna, 2002)

Write the stem clearly and unambiguously. Keep it short and focused on a single idea. Avoid complex sentences, jargon, or double negatives. Do not use “All/None of the above.”

(Haladyna 2002)

(Downing 2005)

Do not use “not,” “except,” or other negative constructions.

Restrictions / conditions

Provide exactly 4 options (one correct + three incorrect). Construct distractors so that they are plausible and attractive to learners who do not know the correct answer, but clearly wrong to those who do.

(Tarrant & Ware, 2010)

Keep all options grammatically parallel, similar in length, and avoid vague terms such as “sometimes” or “the majority.

(Attali & Bar-Hillel, 2003)

Content-validity checking

Provide an excerpt from the source used in the generation of the question and its answers

(Artsi et al., 2024)

Few-shot prompting

Example:

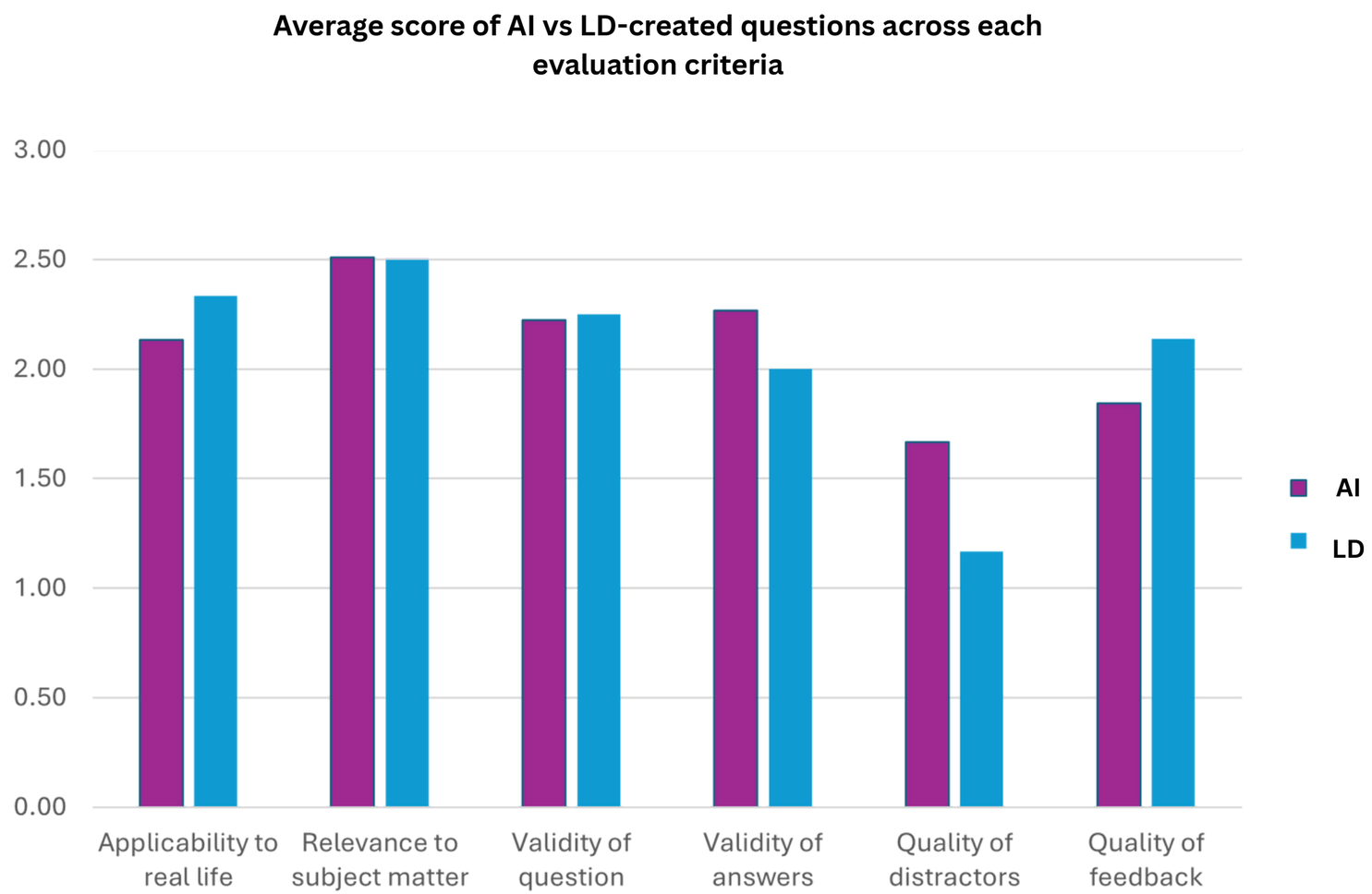
Assessing question quality

- MS Form
- Rubric provided
- Blind study
- All participants had 5+ years of experience creating or assessing questions

Quality criteria	Description
Applicability to real life	Are the questions testing application of skills, or are they recall-based?
Relevance to subject matter	Are the questions relevant to the subject matter?
Validity of question	Are the questions valid and accurate from a content perspective?
Validity of answers	Are the correct answers objectively correct?
Quality of distractors	Are the distractors giving away the correct answer?
Quality of feedback	Is the feedback enough to explain why an answer is correct?

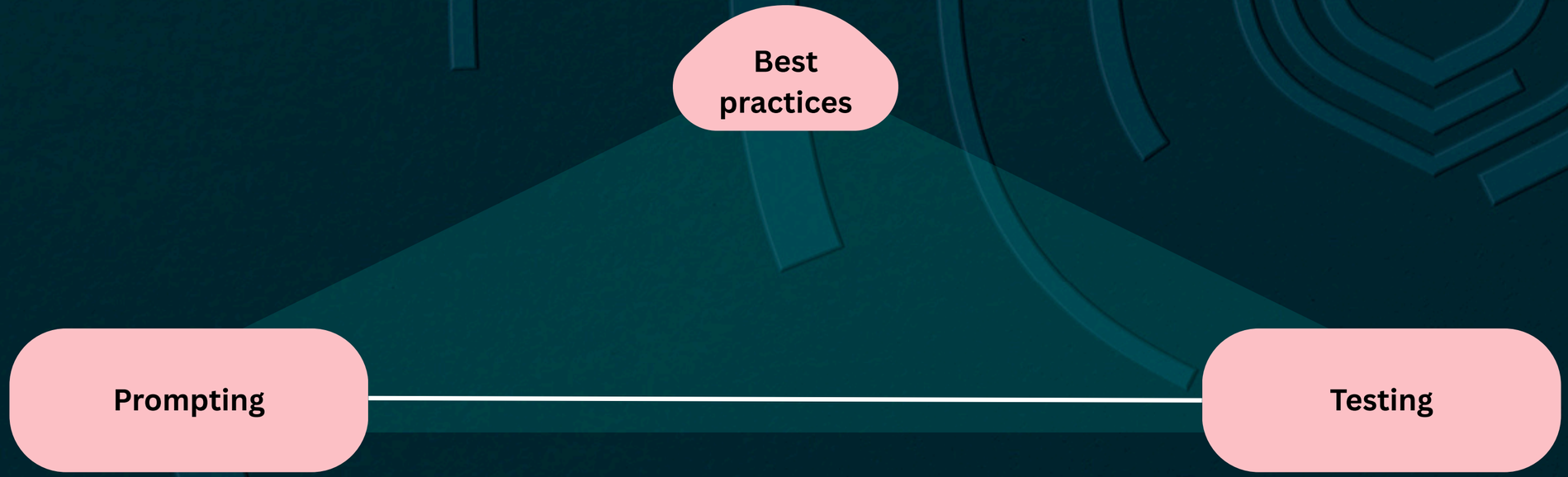
Criterion	1 (Poor)	2 (Fair)	3 (Good)
Applicability to Real Life	The question is purely recall-based and has no real-life application.	The question is mostly recall-based with limited real-life application.	The question tests application of skills and is highly relevant to real life.
Relevance to Subject Matter	The question is irrelevant to the subject matter.	The question is somewhat relevant.	The question is relevant.
Validity of Question	The question is invalid and inaccurate from a content perspective.	The question has some inaccuracies but is mostly valid.	The question is completely valid and accurate.
Validity of Answers	The correct answer is not objectively correct.	The correct answer is mostly correct with minor issues.	The correct answer is completely objectively correct.
Quality of Distractors	The distractors are obvious and give away the correct answer.	The distractors are somewhat obvious but not entirely.	The distractors are highly effective and challenging.
Quality of Feedback	The feedback is insufficient and does not explain why an answer is correct.	The feedback is helpful but could be more detailed.	The feedback is highly detailed and thoroughly explains why an answer is correct.

Assessing question quality



Embedding criticality

Start with an informed idea of what 'good' looks like



...and the humans involved

Thank you.

Contact:
Stewart Utley
Learning Designer, COE
stewart.utley@cambridge.org

References

- Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. Journal of Educational Measurement, 40(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Advances in Health Sciences Education, 10(2), 133–143. https://doi.org/10.1007/s10459-004-4019-5
- Haladyna, T. M. (2002). Developing and validating multiple-choice test items (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Education Today, 30(6), 539–543. Start with an informed idea of what 'good' looks like