A COMPARATIVE STUDY OF ASSESSMENT QUESTION QUALITY: AI-GENERATED VERSUS HUMAN-AUTHORED IN BUSINESS STATISTICS

Tat Kwong YAP* [a] and Eunice ANG [a]

[a] Nanyang Polytechnic, School of IT, Singapore

YAP_Tat_Kwong@nyp.edu.sg*

While generative artificial intelligence (AI) is increasingly used in education, its application to assessment design requires careful evaluation. This paper addresses this need by presenting the findings of a research project that compares the validity and reliability of AI-generated business statistics assessments with those created by subject matter experts (unit leaders and moderators). Our research aims to determine if AI can effectively support the creation of high-quality assessments in this domain. This study compared business statistics assessments created by human experts and generative AI. Two groups of over 100 Diploma in IT students (2023 and 2024 cohorts) served as the control and experimental groups, respectively. The control group completed expert-created assessments, while the experimental group completed AI-generated assessments. Both assessments covered identical business statistics topics, learning outcomes, and difficulty levels. Correlation analysis was used to determine if both assessments measure the same construct. The reliability coefficient Cronbach's alpha was calculated to assess the internal consistency of each assessment. The discrimination index was used to compare the effectiveness of individual questions in differentiating student performance. The results of this study suggest that generative AI can be a valuable tool for assessment creation. The authors found that AI-generated assessments, when guided by expert input, achieve comparable levels of validity and reliability to those developed by subject matter experts. This finding has significant implications for educators, as it demonstrates the potential for efficiently creating diverse and effective assessment questions spanning various statistical topics, difficulty levels, and learning outcomes. This can free up educators' time to focus on other critical aspects of teaching and learning. Generative AI offers a promising avenue for streamlining the assessment question creation process. However, our research demonstrates that the expertise of unit leader and moderator remains crucial for guaranteeing the quality, relevance, and alignment of assessments with learning objectives.

Building upon these findings, the authors will share our experiences and recommendations for using AI prompts to generate business statistics assessment questions, with a focus on achieving appropriate difficulty, question variation, and alignment with learning outcomes. Further research is essential to develop a robust framework, potentially in the form of a playbook, for evaluating and ensuring the quality of AI-generated assessments across different subject areas.

Keywords: *Generative AI Assessment, Assessment Validity and Reliability, Business Statistic Assessment, Co-Creation of Assessment, Comparative Assessment Study.*

Introduction

The rapid advancement of generative (AI) has ushered in a new era of possibilities across various sectors, including education. Among the most intriguing applications is the potential for AI to revolutionize assessment design. However, the integration of AI into such a critical component of learning necessitates rigorous evaluation to ensure the maintenance of quality and validity. This research project tackles this pressing need through a comparative study of assessment question quality, with a specific focus on business statistics, a core unit in the School of IT's freshman curriculum. We aim to investigate whether AI-generated assessments can achieve comparable levels of validity and reliability to those created by experienced subject matter experts. This study was motivated by the desire to explore the potential of AI as a tool to streamline the assessment question creation process, thereby freeing up valuable time for educators to focus on other crucial aspects of teaching and learning.

Literature Review

The use of AI to generate assessment questions is a burgeoning area of research. Several studies have explored AI's potential in creating diverse and challenging assessment items, while also addressing concerns about capturing subject matter nuances and aligning with pedagogical goals. For instance, a study by Owen et al (2023) discusses AI's capability to generate

various test items and emphasizes the necessity for rigorous evaluation to ensure quality and relevance. However, concerns remain regarding AI's ability to accurately capture subject matter nuances and ensure alignment with pedagogical goals. Several journal articles (e.g., Moorhouse, Yeo & Wan, 2023; Karadag, 2023; Xia et al., 2024) discuss these challenges, emphasizing the need for rigorous oversight. Therefore, rigorous evaluation and validation are essential to determine the effectiveness of AI-generated assessments. Additionally, Rezigalla (2024) highlights the importance of validating AI-generated assessments to maintain their reliability and effectiveness. Given the necessity of rigorous validation in AI-generated assessments, researchers have increasingly turned to comparative studies to benchmark AI's effectiveness against human-authored content. Durak, Egin & Onan A (2025) and Law et al (2025) have examined AI-generated essays, code, and other forms of content, often revealing varying degrees of success. These findings highlight the importance of context in determining the appropriateness of AI-generated materials. According to Bowen & Watson (2024), in the domain of assessment, it is critical to determine whether AI can produce questions that rival the quality and effectiveness of those created by experienced educators. Despite AI's ability to generate diverse assessment items, subject matter expertise remains indispensable in ensuring their accuracy and alignment with pedagogical goals. Studies (e.g. Owen et al., 2023; Rezigalla, 2024) emphasize the need for rigorous validation, which underscores the role of human oversight. Ross (2024) further argues that AI should function as an assistive tool rather than replace educators, advocating for a collaborative approach that maximizes both AI efficiency and human expertise.

Methodology

This research project employed a comparative experimental design to investigate the quality of AI-generated versus human-authored business statistics assessments. The methodology was structured to ensure a rigorous and systematic evaluation of validity, reliability, and question discrimination. Two cohorts of Diploma in IT students from the school, each exceeding 100 students, participated in the study. The 2023 cohort served as the control group, while the 2024 cohort served as the experimental group. The study focused on e-assessment of three topics within the business statistics unit, providing a controlled environment to systematically compare assessment outcomes. For the control group of human-authored assessment, subject matter experts, which are the unit leaders and moderators, developed the assessments which were designed to align with the unit's learning outcomes, covering the three specified business statistics topics at predetermined difficulty levels. For the experimental group of AI-generated assessments, Google Gemini was utilized to create assessments. The AI was prompted with specific parameters, including the same business statistics topics, learning outcomes, and difficulty levels as the human-authored assessments. Expert review and guidance were incorporated into the AI-generated assessment creation process to ensure alignment with educational standards. The prompts given to the AI were carefully crafted to try and match the style and content of the human created assessments. Both sets of assessments were meticulously designed to ensure equivalence in terms of content coverage, learning outcomes, and difficulty levels. This was crucial for a valid comparison.

Student performance on both human-authored and AI-generated assessments was collected through the Brightspace Learning Management System (LMS), including detailed data on each question, such as correct and incorrect responses. Correlation analysis was performed to determine the extent to which both assessments measured the same underlying construct. A high correlation coefficient would indicate that both assessments are measuring the same statistical knowledge and skills. Cronbach's alpha was calculated to assess the internal consistency of each assessment. This statistical measure indicates the degree to which the items within each assessment are consistently measuring the same construct. The discrimination index was used to evaluate the effectiveness of individual questions in differentiating student performance. This metric allowed for a comparison of the quality of individual questions generated by AI versus those created by human experts. In addition, the prompts used in generative AI and the modifications made to the AI generated questions by human experts are analysed to identify patterns and best practices.

Results and Findings

The comparative analysis of assessment question quality between AI-generated and human-authored assessments revealed several key findings. Firstly, the correlation analysis of student performance between the two assessment types demonstrated a strong positive correlation (Pearson correlation coefficient $r = 0.9$). This indicates that both assessment methods effectively measured the same underlying construct of business statistics knowledge and skills.

Secondly, the Cronbach's alpha values for both the AI-generated and human-authored assessments were found to be within an acceptable range, indicating a reasonable level of internal consistency. Specifically, the human-authored assessment yielded a Cronbach's alpha of 0.77, while the AI-generated assessment produced a Cronbach's alpha of 0.78. This suggests that both assessment types maintained a similar level of reliability in measuring the targeted learning outcomes.

Thirdly, the discrimination index analysis showed comparable performance between the two assessment types. The average discrimination index for human-authored questions was 0.46, and for AI-generated questions, it was 0.43. While slight variations were observed in individual question discrimination, the overall distribution of discrimination indices across both sets of assessments was similar. Both of the discrimination indices are considered high, indicating that both human-authored and AI-generated questions

effectively differentiated between high- and low-performing students. This suggests that AI-generated questions were as effective as human-authored questions in differentiating between high and low-performing students.

Finally, the analysis of the prompts used for AI generation and the subsequent modifications made by human experts revealed several patterns. Prompts that explicitly included specific learning outcomes, difficulty levels, and examples yielded higher-quality AI-generated questions. Human modifications primarily focused on refining question clarity, addressing minor content inaccuracies, and ensuring alignment with the specific pedagogical approach of the unit.

Discussion

The strong positive correlation between student performance on AI-generated and human-authored assessments indicates that generative AI, when properly guided, can produce assessments that align with the same underlying construct as those created by subject matter experts. This finding aligns with the growing body of research exploring the potential of AI in educational assessment. The comparable Cronbach's alpha values further reinforce the reliability of AI-generated assessments, demonstrating their ability to consistently measure student knowledge and skills. The similar discrimination indices observed across both assessment types suggest that AI-generated questions are capable of effectively differentiating student performance. This is a critical aspect of assessment quality, as it ensures that assessments can accurately identify students who have mastered the learning objectives from those who have not. The slight variations in individual question discrimination highlight the importance of expert review and refinement, even when using advanced AI tools.

The analysis of AI prompts and human modifications provides valuable insights into the effective use of generative AI in assessment creation. The success of prompts that explicitly included learning outcomes and difficulty levels underscores the importance of clear and specific instructions. This aligns with best practices in prompt engineering for educational applications. The human modifications, which focused on refining clarity and accuracy, emphasize the continued need for subject matter expertise in ensuring the quality and relevance of AI-generated assessments.

The findings of this action research have several practical implications for educators and assessment designers. Firstly, generative AI tools like Google Gemini can be effectively utilized to create high-quality assessments, potentially reducing the time and effort required for assessment development. Secondly, the importance of expert review and guidance in the AI-generated assessment process cannot be overstated. While AI can generate questions that align with learning outcomes and difficulty levels, human expertise is essential for ensuring clarity, accuracy, and alignment with pedagogical approaches. However, it is important to acknowledge the limitations of this study. The research was conducted within a specific context, involving Diploma in IT students and focusing on three topics within a business statistics unit. Future research could explore the generalizability of these findings across different disciplines, educational levels, and assessment formats. Additionally, the study focused on the quality of individual assessment questions. Future research could investigate the effectiveness of AI-generated assessments in promoting student learning and engagement.

Conclusions

This action research confirms that generative AI, when guided by expert input, produces assessments with strong correlation, comparable Cronbach's alpha, and similar discrimination indices to human-authored assessments. This underscores AI's potential as a valuable tool in assessment creation, provided it functions in close collaboration with subject matter experts to ensure quality, relevance, and alignment with learning objectives. Moving forward, educators must proactively develop robust frameworks for evaluating AI-generated assessments, ensuring pedagogical principles and student learning remain paramount. This study's findings will directly inform the creation of a practical playbook, empowering educators to confidently integrate and validate AI-driven assessment practices.

References

Bowen, J. A., & Watson, C. E. (2024). *Teaching with AI: A practical guide to a new era of human learning*. Johns Hopkins University Press.

Durak, H.Y, Egin, F & Onan, A (2025). A Comparison of Human-Written Versus AI-Generated Text in Discussions at Educational Settings: Investigating Features for ChatGPT, Gemini and Bing AI. *Wiley Online Library*. https://doi.org/10.1111/ejed.70014

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5, 100151.* https://doi.org/10.1016/j.caeo.2023.100151

Karadag, N. (2023). The impact of artificial intelligence on online assessment: A preliminary review. *Journal of Educational Technology and Online Learning 6(4).* https://www.researchgate.net/publication/374748911_The_impact_of_artificial_intelligence_on_online_assessment_A_preliminary_review

Law, A. Kk., So, J, Lui, C.T, Choi Y.F., Cheung, K.H., Hung, K. KC & Graham, C.A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ. 25(1):208.*

AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination - PubMed

Ng, S.H.S., Chan, H.Y., Wong, J.H.K, Sam, L. & Privitera,, A.J. (2024). A Scoping Review of the Use of Generative AI in Assessment in Higher Education. (PDF) A Scoping Review of the Use of Generative AI in Assessment in Higher Education

Owen, V.S., Abang, K.B., Idika, D.O., Etta, E.O. & Bassey, A.B. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education 19(8), em2307.* Exploring the potential of artificial intelligence tools in educational measurement and assessment - Eurasia Journal of Mathematics, Science and Technology Education

Rezigalla, A.A. (2024). AI in medica education: uses of AI in construction type A MCQ. *BMC Medical Education.24(1):320.* AI in medical education: uses of AI in construction type A MCQs - PubMed

Ross, D (2024). *The Role of Human Expertise in AI-Powered Assessments*. Braide.ai. The Role of Human Expertise in AI-Powered Assessments — Braide.ai

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21, 40.* https://doi.org/10.1186/s41239-024-00468-z