Are AI Test Questions Any Good?
An Evaluation of Student Performance on AI vs. Human-Generated Exam Items

Carolyne B. Savage
East Texas A&M University
Carolyne.Savage@etamu.edu

Karen Wright
East Texas A&M University

Robert L. Williams
East Texas A&M University
Bob.Williams@etamu.edu

Are AI Test Questions Any Good? An Evaluation of Student Performance on AI vs. Human-Generated Exam Items

Introduction

Artificial Intelligence (AI) has captivated business, research, and the public alike due to its potential in reshaping how we interact with these fields (Sætra, 2023). This technology is purported to improve efficiency and reduce burdens on workers, particularly in the fields of healthcare, education, and agriculture (Pehrson, 2025, Mensah and Swortzel, 2026). Brynjolfsson et al. (2025) report a 15% increase in worker productivity utilizing AI technology, and large gains in its ability to solve problems that human workers have little baseline experience with. Mensah et al. (2026) found that Extension educators utilized AI to support administrative writing tasks, research idea generation, educational programming, and answering clientele questions.

AI is particularly important in education, as teachers are tasked with preparing students for a workforce demanding fluency with this technology (Pehrson, 2025). Various studies evaluate AI in education, but the bulk of research focuses on its use in language learning (17%), computer science (16%), and engineering (12%) (Crompton and Burke, 2023). High school educators report using AI primarily in administrative tasks like creating assignments, generating ideas, and developing instructional materials (Pehrson et al., 2025). In higher education, AI is used primarily in understanding and promoting student performance through assessment, tutoring, and predictive tools (Crompton and Burke, 2023). The author found no studies showing AI's use in creating tests. According to Pehrson et al. (2025), there is a lack of knowledge regarding AI's integration into agricultural education specifically. Extension is most frequently researched, but only four recently published studies were found on this topic (Mensah et al., 2026, Mensah and Swortzel, 2026, Priya et al., 2025, Khatri et al., 2024).

The objective of this study is to evaluate the feasibility of utilizing AI as a test-building tool for the Sustainable Agriculture Career Development Event (CDE) hosted at East Texas A&M by assessing participant performance and testing the reliability of AI-generated questions. This study integrates into the national agricultural research value of increasing prosperity through innovation in AFNR systems (AAAE, 2023). By advancing the use of AI in agricultural education, educators can increase workflow, reduce administrative burdens, and focus more effectively on their own educational priorities.

Methodology

Twenty-five human-generated and 25 AI-generated questions were used for the 2026 Sustainable Agriculture CDE. Primary sources were uploaded into ChatGPT and prompted to generate 25 multiple-choice questions from each source. Over 200 questions were created from the input sources. To select 25 for use in this study, an allocative proportion formula was used to determine how many questions from each source were required to appropriately represent the content on the exam. Chat GPT was used to determine and generate this formula, as well as randomly select the number of test questions to be used. Calculations were performed by hand.

The 50-question written exam was administered to 16 participants. Questions 1-25 (Exam 1) were human-generated, and questions 26-50 (Exam 2) were AI-generated. Results, including mean, standard deviation, and section scores, were obtained through publicly available information on JudgingCard.com. Data pertaining to individual questions was hand coded in

Microsoft Excel (2024), with correct answers denoted by a “1”, and incorrect answered denoted by a “0”.

Participant scores were evaluated using split-half reliability and Spearman-Brown coefficient in Microsoft Excel (2024) to determine the internal consistency of test questions. Additionally, a paired-t test was undertaken in Statistical Analysis Software 9.4 to evaluate the mean scores differences between Exam 1 and Exam 2.

Results and Conclusions

The average score for Exam 1 was 22.88, with a minimum of 14, a maximum of 34, and a standard deviation of 6.69. The average score for Exam 2 was 33, with a minimum of 18, maximum of 42, and a standard deviation of 7.27. No participant received a perfect score (50 points) on either section. This indicates that participants performed better with AI-generated questions (Exam 2) than human-generated questions (Exam 1). Participants on the whole scored lower on Exam 1, but the variation in their scores was smaller.

T-test results indicate statistically significant differences in test performance on Exam 2 compared to Exam 1 (t-statistic=7.38, $Pr < 0.0001$). The Spearman-Brown coefficient of 0.82 indicates a high degree of reliability between Exam 1 and Exam 2. Participants’ overall performance and t-test results indicate that AI-generated test questions were, on average, “easier” than human-generated questions. However, the split-half reliability results indicate that both human and AI-generated questions reliably tested the concepts of the exam. Therefore, this study concludes that AI shows potential as an exam-building tool.

Implications

AI presents opportunities for educators to increase productivity, reduce administrative burdens, and enhance student outcomes. However, the utilization of AI in higher education has largely focused on supporting learners in their educational ventures. Less is known about the feasibility of integrating AI into curriculum development and assessment. The objective of this study was to evaluate the reliability of and participants' performance on AI-generated test questions.

Further research should explore additional AI-generated test questions to evaluate the limits of difficulty that can be achieved with this technology. Future researchers should ask, “how far can we stretch AI to reasonably challenge our students?” Beyond strict performance-based evaluations, an investigation of the language used in AI-generated questions should be undertaken. Future researchers should ask, “*why* are students performing the way they do on AI-generated questions?” The world’s knowledge is at our fingertips unlike any other point in history, ready to be reshuffled and propagated anew with a single command. By utilizing this innovation, prosperity in AFNR systems can be increased through educating students beyond the limits of personal knowledge expertise and streamlining workflow to advance educators' research and instructional goals.

References

American Association for Agricultural Education (AAAE). (2023). *AAAE Research Values*.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889-942. https://doi.org/10.1093/qje/qjae044

Crompton, H., Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Education Technology in Higher Education, 20*(22). https://doi/org/10.1186.s41239-023-00392-8.

Khatri, A., Lallawmkimi M.C., Panigrahi, C.K., Minj, A., & Koushal, S., Ali, M. U. (2024). Integration of ICT in Agricultural Extension Services: A review. *Journal of Experimental Agricultural International, 26*(12)394-410. https://doi.org/10.9734/jeai/2024/v46i123146

Mensah, E. A., Swortzel, K. A., Osman, N. (2026) Towards the integration of Generative AI into the Cooperative Extension Service: Insights for guiding professional development. *Journal of Agricultural Education, 67*(2). https://doi.org/10.5032/jae.v67i2.3302

Mensah, E. A., Swortzel, K. A. (2026). Generative AI and Extension program planning and evaluation: Do general large language models answer domain-specific questions accurately? *Journal of Agricultural Education, 67*(1). https://doi.org/10.5032/jae.v67i1.3289.

Pehrson, S., Sorensen, T. J., Wood, M. J., Burrows, M. S., & Furse, J. (2025). Examining AI perception and utilization among agricultural educators in Utah. *Journal of Agricultural Education, 66*(3). https://doi.org/10.5032/jae.v66i3.2910.

Priya, N.K., Patil, S. S., Mishra, I., Tripathi, S., Pathak, A. K., & Shastri, S. (2025). Digital innovation in extension education: A review of emerging technologies and their impact on agricultural knowledge dissemination. *Journal of Experimental Agriculture International, 47*(5)408-423. https://doi.org/10.9734/jeai/2025/v47i53430.

SirGil (2021, March 5). *Split-half Method (Reliability Analysis)*. [Video]. YouTube. https://www.youtube.com/watch?v=xfzBjHHXcZU&t=329s