

Are AI Test Questions Any Good?

An Evaluation of Student Performance on AI vs. Human-Generated Exam Items

Carolyn B. Savage, Karen Wright, Robert L. Williams



Scan to find my LinkedIn!

Introduction

- Artificial Intelligence (AI) may improve efficiency and reduce burdens on workers, particularly in agriculture and education (Pehrson et al. 2025, and Mensah & Swortzel, 2026).
- Teachers are tasked with training pupils in utilizing new technology, making AI skills vital to education (Pehrson, 2025).
- Educators commonly use AI for administrative tasks (Pehrson et al. 2025).
- There is a lack of knowledge about AI's use in agriculture education (Pehrson et al. 2025).
- Little is known about the feasibility of using AI for test building.
- This study integrated in the national agricultural research value of increasing prosperity through innovation in AFNR systems (AAAE, 2023).

Methodology

Exam Item Creation: Source material was uploaded to ChatGPT and promoted to generate 25 multiple-choice exam items from each source. Over 200 items were created. An allocative proportional formula was used to determine how many questions from each source to put in the exam.

- Questions should be appropriate for high school FFA students
- Questions should include four answer choices (A-D)
- Questions should be clear and competition-style
- Include a combination of easy, moderate, and difficult questions
- Use exact labeled PDF page numbers
- Avoid overly complex wording

Figure 1. Prompts to generate exam items.

Test Administration: 25 human-generated and 25 AI-generated multiple-choice exam items were utilized in the 2026 Sustainable Agriculture Career Development Event at East Texas A&M University.

Questions 1-25 (Exam 1) were human-generated, and questions 26-50 (Exam 2) were AI-generated. Data were obtained through publicly available information on JudgingCard.com.

Data Coding: Data were hand-coded in Microsoft Excel (2024), with correct answers denoted by "1" and incorrect answers denoted by "0".

Statistical Analysis: Split-half reliability and Spearman-Brown coefficient were calculated in Microsoft Excel (2024) to determine internal consistency of exam items.

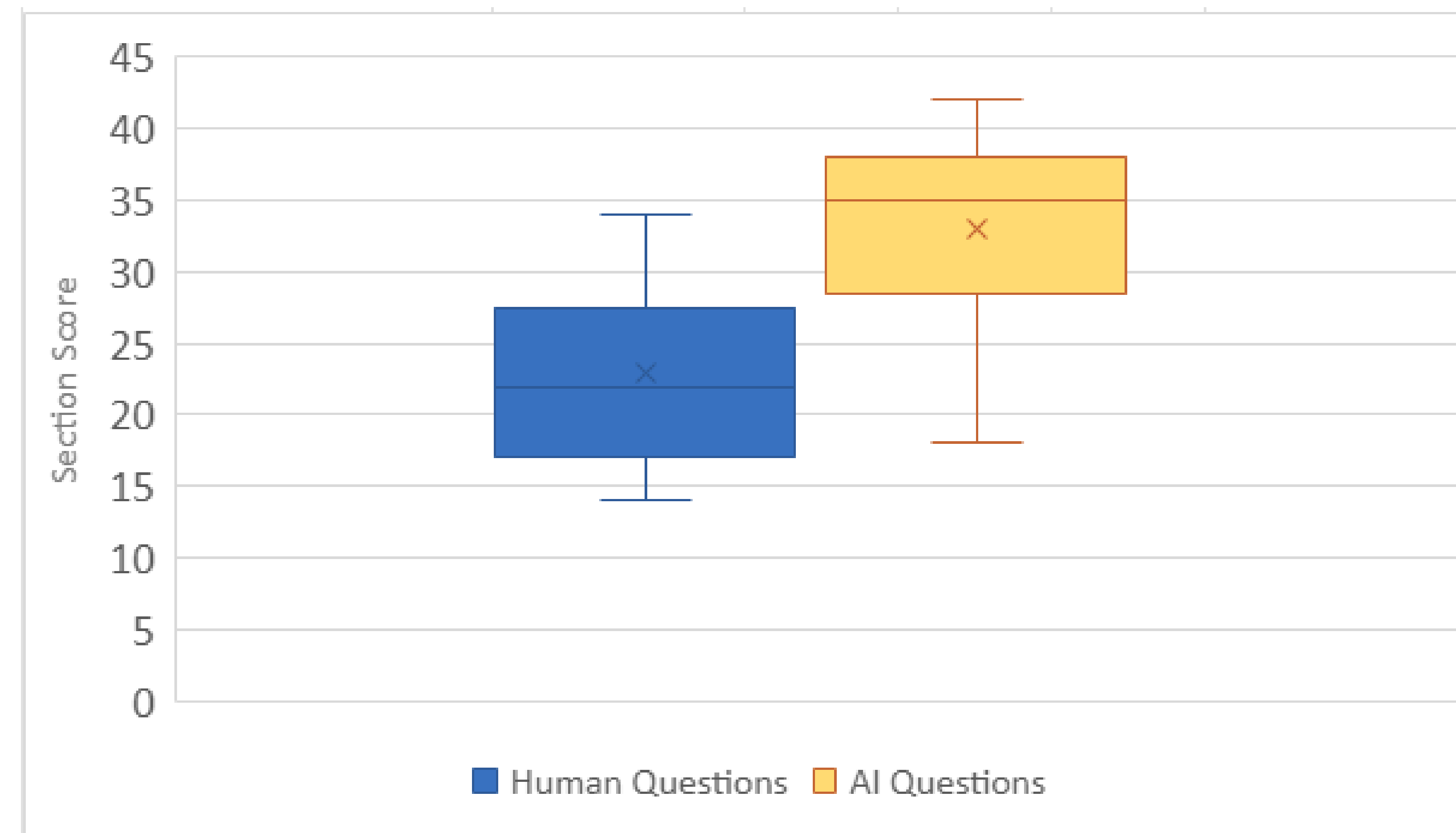


Figure 2. Boxplots by exam item source.

Item	Multiple Choice Answer Option			
	A	B	C	D
1	9	4	0	3
2	12	1	2	1
3	2	3	11	0
4	1	12	1	2
5	0	11	5	0
6	10	2	2	2
7	2	0	13	1
8	2	4	3	7
9	4	2	8	2
10	6	4	3	3
11	3	11	1	1
12	0	12	3	1
13	5	0	0	11
14	2	3	11	0
15	10	2	3	1
16	2	12	1	1
17	1	5	0	10
18	1	13	2	0
19	1	3	3	9
20	6	1	5	4
21	3	3	3	7
22	8	3	1	4
23	10	0	2	4
24	4	5	2	5
25	2	5	6	3

Item	Multiple Choice Answer Option			
	A	B	C	D
26	0	0	16	0
27	0	7	7	2
28	1	1	10	4
29	0	0	0	16
30	14	1	0	1
31	1	14	1	0
32	1	15	0	0
33	6	0	9	1
34	2	1	12	1
35	1	1	14	0
36	4	6	1	5
37	12	1	1	2
38	6	10	0	0
39	14	0	2	0
40	0	7	1	8
41	10	1	2	3
42	4	6	1	3
43	2	8	2	4
44	3	12	1	0
45	6	3	5	2
46	1	1	13	1
47	1	13	0	2
48	2	10	1	3
49	1	10	3	2
50	2	11	1	2

Table 1. Exam item analysis. Number of responses per multiple choice answer option inside table. Correct answer option bolded.

Results and Conclusions

- Average Exam 1 (human-generated) score was 22.88/50 (min 14, max 34, std. 6.69).
- Average Exam 2 (AI-generated) score was 33/50 (min 18, max 42, std. 7.27).
- Exam 1 had lower scores overall, but with less variation.
- T-test indicated statistically significant differences between Exam 1 and Exam 2 scores.
- Spearman-Brown coefficient=0.82.
- Both human and AI-generated question reliably tested exam concepts.
- AI-generated questions were generally "easier" than human-generated.
- AI shows promise as a test building tool.

Implications

- AI is a powerful tool for increasing teacher productivity and enhancing student outcomes.
- AI for test building could increase interdisciplinary educational opportunities for students, regardless of educator expertise.
- Potential future research questions:
 - "How far can we stretch AI to reasonable challenge our students?"
 - "Why are students performing how they do on AI-generated questions?"

- Include a combination of easy, moderate, and difficult questions
- Make these easier for introductory FFA students
- Create "B" versions of difficult questions that are more simple, but cover the same content
- Create a separate answer key
- Clear spacing and consistent formatting
- APA-style references

Figure 3. Prompts to generate exam items.

Acknowledgements

This project was partially funded through the USDA NIFA NLGCA Award no. 2023-70001-40987.



Scan to access the full abstract!

References

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