Vol. 08, No. 1 (2026) 591-596, doi: 10.24874/PES08.01B.024

Proceedings on Engineering Sciences

www.pesjournal.net

CAN CHATGPT-GENERATED EXAMS USING BLOOM'S TAXONOMY EFFECTIVELY ASSESS COGNITIVE LEARNING OUTCOMES IN ROAD ENGINEERING EDUCATION?

Yasmany García Ramírez[1]

Received 23.04.2025.
Revised 17.07.2025.
Accepted 05.08.2025.

Keywords:

Bloom's Taxonomy, Road Design Learning, Cognitive Assessment

A B S T R A C T

This study investigates the effectiveness of ChatGPT-generated multiple-choice exams in evaluating cognitive learning outcomes in civil engineering education, specifically in the subject Road Construction I at Universidad Técnica Particular de Loja, Ecuador. Using the revised Bloom's Taxonomy as a framework, a 32-question exam was developed, covering the first four cognitive levels: remember, understand, apply, and analyze. The test was administered to 101 students divided into two groups, and the results were analyzed based on difficulty and discrimination indices, as well as internal reliability using the KR-20 coefficient. Findings indicate that while ChatGPT-generated questions demonstrated acceptable internal reliability (KR-20 > 0.7) and discrimination indices, but reveal that 40–50% of questions fell outside the optimal difficulty range. Unexpectedly, higher-order cognitive questions yielded better scores, underscoring both the potential and challenges of AI in creating balanced assessment. This study underscores the potential of ChatGPT as a tool for generating assessment instruments but also identifies limitations, particularly in creating balanced difficulty distributions and higher-order cognitive questions.

1. INTRODUCTION

Advances in artificial intelligence (AI) have transformed educational assessment, offering tools like AI language models (AI LMs) to streamline the creation of multiple-choice questions (MCQs) and other evaluation materials (Alves de Castro, 2023; Perez Sanpablo et al., 2024). Models such as ChatGPT have demonstrated their capacity to produce high-quality MCQs efficiently, with recent iterations, like GPT-4, achieving improved reliability and alignment with structured learning objectives (Cheung et al., 2023; Herrmann-Werner et al., 2024).

At the same time, Bloom's Taxonomy has provided a foundational framework for designing assessments that evaluate cognitive processes, from basic recall to advanced problem-solving (Anderson & Krathwohl, 2001; Bloom, 1956). Widely adopted across disciplines, this taxonomy ensures assessments support structured

[1] Corresponding author: Yasmany García-Ramírez
Email: ydgarcia1@utpl.edu.ec

learning, fostering both foundational knowledge and critical thinking (Dorodchi et al., 2017).

Combining these two approaches—AI-driven assessment tools and structured cognitive frameworks—opens promising avenues for innovation in educational evaluation. However, the development of balanced assessments that accurately reflect cognitive complexity remains a persistent challenge, particularly in specialized and technical disciplines.

Despite the growing use of AI tools in education, limited research has examined their application in domain-specific contexts such as civil engineering, and even less so in foundational subjects like road engineering. . This study seeks to address that gap by evaluating the reliability, validity, and cognitive alignment of ChatGPT-generated exams, structured according to Bloom's Taxonomy, within the context of the Road Construction I course at Universidad Técnica Particular de Loja (Ecuador).

In doing so, it aims to assess whether AI-generated assessments can effectively measure learning outcomes in a structured and pedagogically sound manner. By focusing on a specific engineering context, the study contributes to the broader discussion on integrating AI in higher education assessment while identifying both the potential and limitations of current language model capabilities.

2. MATERIALS AND METHODS

2.1 Participants

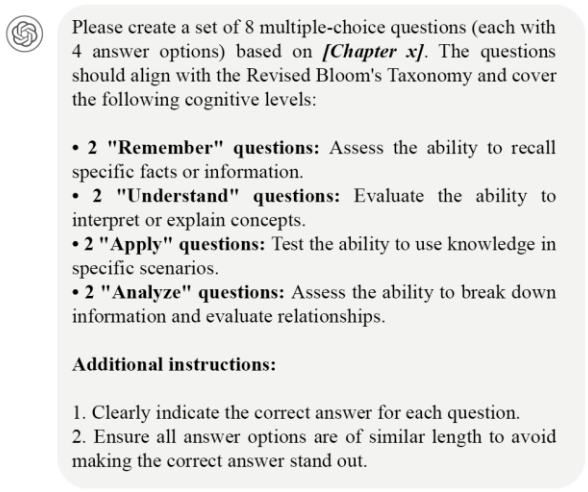
The study involved 101 students enrolled in the Road Construction I course, divided into two groups: Group A (51 students) and Group B (50 students).

2.2 Test design

A 32-question multiple-choice test was created using the limited free plan of ChatGPT-4 (accessed on October 20, 2024, via https://openai.com/index/openai-api/). Each question was designed to correspond to one of the first four cognitive levels of the Revised Bloom's Taxonomy (refer to Figure 1):

- Remember (8 questions): Focused on recalling factual information.
- Understand (8 questions): Assessed interpretation and explanation of concepts.
- Apply (8 questions): Evaluated the application of knowledge in context-specific scenarios.
- Analyze (8 questions): Tested the ability to break down and evaluate relationships.

The book "Design of Geometric and Operation of Two-Lane Highways" (García-Ramírez, 2022) served as the foundation for the exams, outlining the key topics to be evaluated. These included fundamental areas such as road introduction, driving principles, traffic studies, and route analysis, ensuring the assessments effectively measure comprehensive student learning outcomes.


Please create a set of 8 multiple-choice questions (each with 4 answer options) based on ***[Chapter x]***. The questions should align with the Revised Bloom's Taxonomy and cover the following cognitive levels:

- **2 "Remember" questions:** Assess the ability to recall specific facts or information.
- **2 "Understand" questions:** Evaluate the ability to interpret or explain concepts.
- **2 "Apply" questions:** Test the ability to use knowledge in specific scenarios.
- **2 "Analyze" questions:** Assess the ability to break down information and evaluate relationships.

Additional instructions:

1. Clearly indicate the correct answer for each question.
2. Ensure all answer options are of similar length to avoid making the correct answer stand out.

Figure 1. ChatGPT prompt used to generate exam questions for this study

2.3 Procedure

The test was administered under timed conditions (30 minutes). After completion, the responses were graded, and the following analyses were conducted:

- KR-20 reliability: Overall reliability of the test, measuring internal consistency.
- Difficulty index: Proportion of students who answered correctly per question.
- Discrimination index: Ability of questions to differentiate between high- and low-performing students.
- Cognitive performance trends: Average scores by cognitive level were analyzed.

2.4 Data analysis

Data was analyzed statistically to compare group performance and evaluate the reliability and validity of the test. Graphical and tabular representations were used to interpret trends.

3. RESULTS

3.1 Evaluating answer length ratios for test validity

The initial analysis involved counting the number of characters in each answer option. The length of the correct answers was then compared to the average length of the incorrect answers, and the resulting ratio is shown in Figure 2. The analysis considered four topics and both exam versions (Groups A and B). Ideally, this ratio should be close to 1, ensuring students cannot guess the correct answer based solely on its length.

Figure 2 demonstrates that while most questions in both versions maintain a ratio near 1, several exceed this threshold, with some slightly surpassing a ratio of 1.5. A linear regression analysis was conducted between this ratio and the difficulty index, revealing no significant relationship. This indicates that the length of the correct answer did not influence students' choices.

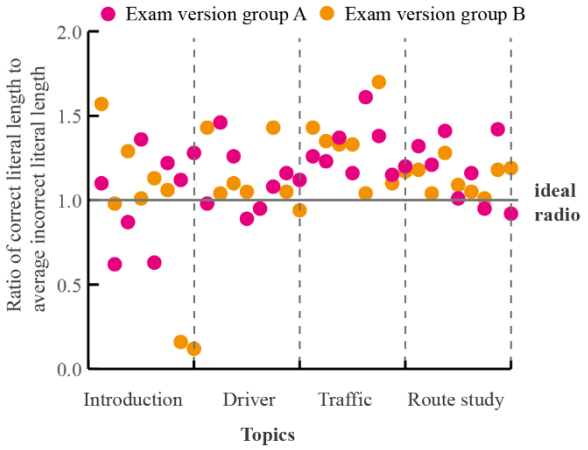


Figure 2. Correct-to-incorrect answer length ratios across topics and exam versions.

3.2 Kuder-Richardson Coefficient (KR20)

The The KR20 analysis (Kuder & Richardson, 1937) analysis was conducted to assess the internal reliability of the exam. Student responses were coded as 1 for correct answers and 0 for incorrect answers.

For Group A, the Σpq was 5.80, and the variance was 20.20, resulting in an KR20 value of 0.74, which is considered acceptable. For Group B, the Σpq was 5.99, and the variance was 20.10, resulting in an KR20 value of 0.73, also deemed acceptable. These results suggest that both exam versions possess acceptable internal reliability and are consistent in measuring students' knowledge of road geometric design.

3.3 Difficulty and discrimination indices

The difficulty and discrimination indices were calculated to evaluate the quality of the questions and their ability to differentiate between high- and low-performing students. To this end, the top 27% and bottom 27% of students, based on scores, were analyzed.

The difficulty index ranged from 0.25 to 0.85. A value near 0 indicates a very difficult question, while a value close to 1 reflects an easy question. The optimal range is between 0.3 and 0.7, representing moderate difficulty conducive to evaluating learning outcomes. For Group A, 40.6% of the questions fell within the optimal range, 53.1% were relatively easy, and 6.3% were considered difficult. In Group B, 37.5% of questions were in the optimal range, 46.9% were easy, and 15.6% were difficult. Group A's difficulty index ranged from 0.10 to 0.90, while Group B's ranged from 0.20 to 1.00. This indicates room for improvement in these items and their response options.

In Figure 3, regression lines were added to observe trends relative to Bloom's taxonomy levels. The first two questions for each topic focused on "Remember," while the final questions targeted "Analyze." Ideally, these lines should indicate easier questions at the start and more difficult questions at higher Bloom levels. However, most regression lines showed an ascending trend, contrary to expectations.

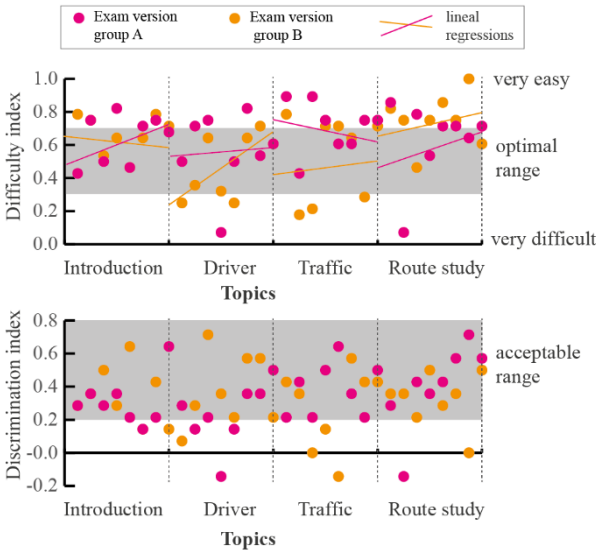


Figure 3. Difficulty index trends across Bloom's taxonomy levels and exam groups.

The discrimination index, also included in Figure 3, ranged from -1 to 1. A positive value indicates that a question effectively differentiates between high- and low-performing students, with values above 0.2 being most desirable. In Group A, 84.38% of questions exceeded the 0.2 threshold, while in Group B, 78.3% achieved this value. Despite these high percentages, some items could require revision. Specifically, 15.63% of questions in Group A and 21.8% in Group B should be discarded due to discrimination indices below 0.2.

3.4 Cognitive performance trends

Finally, the students' scores in both groups were analyzed concerning Bloom's taxonomy levels. Table 1 shows the average scores for each level.

Table 1. Average scores by Bloom's taxonomy level for both groups.

Bloom taxonomy level	Group A (Score/10)	Group B (Score/10)
Remember	5.8	5.8
Understand	6.3	5.4
Apply	6.6	6.5
Analyze	7.1	6.7

Interestingly in Table 1, the lowest cognitive levels yielded the lowest scores, while higher cognitive levels produced higher scores. This contrasts with previous studies, where students typically scored higher on lower cognitive levels, with scores decreasing as cognitive complexity increased.

4. DISCUSSION

This study contributes to understanding the effectiveness of AI-generated exams in evaluating cognitive learning outcomes in road geometric design education, with a focus on metrics such as difficulty and discrimination indices, internal reliability, and cognitive performance trends.

The exam design, based on Bloom's Taxonomy, aimed to evaluate cognitive processes ranging from basic recall to complex analysis. Similar to findings by (Bajčetić et al., 2024), which highlighted the challenges AI faces in classifying questions into Bloom's levels, this study observed unexpected trends in the difficulty distribution. While an optimal balance was targeted, 40.6% of the questions in Group A and 37.5% in Group B fell within the ideal range of difficulty. However, a significant proportion were either too easy or too difficult, suggesting a need to refine question design for better alignment with cognitive assessment goals. Previous study noted that AI often struggles with creating balanced difficulty across cognitive levels, especially at higher-order levels (Crowther et al., 2023).

The discrimination index showed that most questions effectively distinguished between high- and low-performing students, with 84.38% in Group A and 78.3% in Group B exceeding the desirable threshold of 0.2. However, a small percentage of questions failed to meet this criterion, necessitating revision or elimination of these items. These findings are consistent with (Govender, 2024), who noted that AI-generated questions often need refinement to achieve better discriminatory power at higher cognitive levels.

The KR20 reliability coefficients for both groups (0.74 for Group A and 0.73 for Group B) indicate acceptable internal consistency, corroborating studies like (Cheung et al., 2023), which reported similar reliability levels in AI-assisted assessments. However, the presence of an excess of easy questions and inconsistent difficulty patterns across cognitive levels undermines the test's ability to comprehensively evaluate learning outcomes. Previous research emphasized the importance of addressing such issues in AI-generated assessments to enhance validity and reliability (Farazouli et al., 2024).

Performance trends across Bloom's levels revealed that students achieved higher average scores in higher-order cognitive levels (Apply and Analyze) compared to lower levels (Remember and Understand). This is contrary to prior studies (e.g., (Bajčetić et al., 2024; Crowther et al., 2023)), which generally observed declining performance as cognitive demands increased. The upward trend in scores might reflect the instructional strategies employed but also suggests potential misalignment between test items and Bloom's levels, warranting further investigation.

Finally, while the use of AI in exam generation shows promise, challenges remain in ensuring optimal question difficulty, precise alignment with cognitive levels, and the avoidance of unintentional cues such as correct answer length. The findings echo concerns by (Elsayed, 2023), who proposed optimization algorithms to enhance question quality, particularly for higher-order thinking skills.

This study has several limitations. First, the reliance on AI-generated exams introduces potential biases in question formulation. Additionally, the limited sample size and focus on a single subject area (road geometric design) restrict the generalizability of the findings. Future research should explore larger samples, interdisciplinary applications, and alternative assessment formats, such as open-ended or scenario-based questions, to address these limitations and enhance the overall validity of AI-generated exams in education.

5. CONCLUSION

The study demonstrates that ChatGPT holds promise as a tool for generating multiple-choice exams in engineering education, offering efficiency and coherence in question design. However, its application reveals significant challenges that require attention to fully harness its potential. While the generated tests achieved acceptable levels of reliability and discrimination, the difficulty index results suggest an imbalance, with many questions failing to align with optimal difficulty levels. This imbalance compromises the exam's capacity to comprehensively assess the students' cognitive abilities. A notable finding was the unexpected trend in cognitive performance, where higher-order cognitive questions, as defined by Bloom's Taxonomy, yielded better student performance than lower-level ones. This deviation from established assessment patterns highlights the complexities of aligning AI-generated questions with cognitive expectations and suggests a need for careful calibration of difficulty and content relevance. The results underline the value of combining AI-generated assessments with human oversight to enhance question quality, maintain validity, and align with educational objectives. While ChatGPT has proven to be a valuable resource in streamlining assessment creation, its limitations suggest that it should complement rather than replace traditional test design practices. Future research should explore optimizing AI's capabilities to support the creation of more sophisticated and balanced assessments across cognitive levels in engineering education.

References:

Alves de Castro, C. (2023). A Discussion about the Impact of ChatGPT in Education: Benefits and Concerns. *Journal of Business Theory and Practice*, *11*(2), 28–34. https://doi.org/10.22158/JBTP.V11N2P28

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman.

Bajčetić, M., Mirčić, A., Rakočević, J., Đoković, D., Milutinović, K., & Zaletel, I. (2024). Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. *Annals of Anatomy - Anatomischer Anzeiger*, *254*(Jun), 1–7. https://doi.org/10.1016/J.AANAT.2024.152261

Bloom, B. S. (1956). *Taxonomy of Educational Objectives* (Ann Arbor). Edwards Brothers.

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *Plos One*, *18*(8), 1–12. https://doi.org/10.1371/JOURNAL.PONE.0290691

Crowther, G. J., Sankar, U., Knight, L. S., Myers, D. L., Patton, K. T., Jenkins, L. D., & Knight, T. A. (2023). Chatbot responses suggest that hypothetical biology questions are harder than realistic ones. *Journal of Microbiology & Biology Education*, *24*(3). https://doi.org/10.1128/jmbe.00153-23

Dorodchi, M., Dehbozorgi, N., & Frevert, T. K. (2017). "I wish I could rank my exam's challenge level!": An Algorithm of Bloom's Taxonomy in teaching CS1. *Proceedings - Frontiers in Education Conference, FIE*, *2017-October*, 1–5. https://doi.org/10.1109/FIE.2017.8190523

Elsayed, S. (2023). Towards Mitigating ChatGPT's Negative Impact on Education: Optimizing Question Design Through Bloom's Taxonomy. *IEEE Region 10 Symposium (TENSYMP)*, 1–6. https://doi.org/10.1109/TENSYMP55890.2023.10223662

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, *49*(3), 363–375. https://doi.org/10.1080/02602938.2023.2241676

García-Ramírez, Y. (2022). *Diseño geométrico y operación de carreteras de dos carriles* (1st ed.). Ediciones de la U.

Govender, R. G. (2024). My AI students: Evaluating the proficiency of three AI chatbots in completeness and accuracy. *Contemporary Educational Technology*, *16*(2), 1–13. https://doi.org/10.30935/cedtech/14564

Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., Zipfel, S., & Mahling, M. (2024). Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. *Journal of Medical Internet Research*, *26*(1), e52113. https://doi.org/10.2196/52113

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Perez Sanpablo, A. I., Arquer Ruiz, M. del C., Meneses Peñaloza, A., Rodriguez Reyes, G., Quiñones Uriostegui, I., & Anaya Campos, L. E. (2024). *Development and Evaluation of a Diagnostic Exam for Undergraduate Biomedical Engineering Students Using GPT Language Model-Based Virtual Agents* (J. d. J. A. Flores Cuautle, Ed.; pp. 128–136). https://doi.org/10.1007/978-3-031-46933-6_14

Yasmany García-Ramírez
Universidad Técnica Particular de Loja,
Loja,
Ecuador
ydgarcia1@utpl.edu.ec
ORCID 0000-0002-0250-5155