Randwick International of Education and Linguistics Science (RIELS) Journal
Vol. 4, No. 4, December 2023 | Page 937-944 |
https://www.randwickresearch.com/index.php/rielsj

Examining the Validity and Reliability of ChatGPT 3.5-Generated Reading Comprehension Questions for Academic Texts

DOI: *https://doi.org/10.47175/rielsj.v4i4.835*

| Meida Rabia Sihite[1] | Meisuri[2] | Berlin Sibarani[3] |

[1] *English Education Study Program, Universitas Alwashliyah, Medan, Indonesia*

[2,3] *Postgraduate Program, Universitas Negeri Medan, Medan, Indonesia*

[1]*meidarabia55@gmail.com,*
[2]*meisuri@yahoo.com,*
[3]*berlinsibarani@unimed.ac.id*

ABSTRACT

This research examines the capacity of ChatGPT 3.5 in generating reading comprehension questions for academic texts, with a focus on their alignment with higher-order cognitive skills as per Bloom's Taxonomy. A paper-based test comprising 30 multiple-choice questions was constructed using ChatGPT 3.5, based on three selected TOEFL ITP reading comprehension passages. The study employed a mixed-methods approach, integrating qualitative content analysis to assess the cognitive level of each question and quantitative methods to analyze student responses. Data collection involved administering the AI-generated questions to students and scoring their responses. Analysis techniques included Pearson correlation coefficients to determine validity and reliability analysis using Cronbach's Alpha to measure internal consistency. The findings revealed that ChatGPT 3.5 is capable of producing questions that cover a range of cognitive levels, from analysis to creation, however only 10 out of 30 questions met the validity criteria, indicating a need for improvement in the AI's question generation process. The reliability of these questions was moderate, suggesting a reasonable level of internal consistency. The study concludes that while AI-generated questions show promise in educational assessments, ongoing improvement of AI models is necessary to enhance their effectiveness. The implications of this research are significant for the future integration of AI in educational settings, indicating a potential role for AI in developing meaningful assessment tools. The study recommends future research to explore various question types and incorporate student feedback to optimize the effectiveness of AI in education.

KEYWORDS

ChatGPT 3.5; validity; reliability; reading comprehension questions

INTRODUCTION

The integration of AI language models, such as ChatGPT 3.5, in educational and language education domains has garnered significant attention due to their potential to revolutionize the creation of educational materials and language assessments. These advanced models have demonstrated utility in generating virtual patient simulations, quizzes, and educational materials, showcasing their potential to enhance learning experiences (Eysenbach, 2023; Hung et al., 2023). However, concerns have been raised regarding the reliability and validity of AI-generated content, especially when applied to assess reading comprehension proficiency in academic texts (Rahman & Watanobe, 2023; Tyson, 2023; Vandiver, 2008) .

The existing literature has highlighted the potential of AI in education, with studies demonstrating the increasing utility of AI as a national growth engine and its potential to provide tremendous value across various educational fields (Su & Yang, 2023). The use of artificial intelligence (AI) in education has gained significant attention in recent years, with a growing number of educational institutions and organizations exploring the potential benefits of AI-driven technologies (Su & Yang, 2023). However, the specific application of AI in generating reading comprehension questions for academic texts and its impact on accurately measuring students' reading comprehension proficiency remains an area that requires further investigation.

While AI models have shown potential in creating educational materials, there are associated risks, such as the potential diminishment of critical thinking skills and educational inequalities (Schiff, 2021). This emphasizes the importance of critically evaluating the validity and reliability of ChatGPT 3.5-generated reading comprehension questions for academic texts. The development and implementation of AI-based interventions aimed at promoting language proficiency and addressing literacy challenges among students with diverse linguistic backgrounds have been identified as crucial areas for exploration (Cukurova et al., 2019). This highlights the need to assess the effectiveness of AI-generated language tests and explore the potential of AI-generated items in language assessment.

The existing literature has provided valuable insights into the potential applications of AI in education, particularly in language education and assessment. However, there is a notable gap in the literature regarding the specific examination of the reliability and validity of ChatGPT 3.5-generated reading comprehension questions for academic texts. Therefore, conducting this research is crucial to address this gap and provide evidence-based insights into the effectiveness of AI-generated reading comprehension questions in accurately measuring students' reading comprehension proficiency in the context of academic texts.

RESEARCH METHODS

Research Design

The research employed a mixed method of qualitative and quantitative approaches to comprehensively evaluate the reliability and validity of ChatGPT-generated reading comprehension questions for academic texts (Kim, 2015). The qualitative approach was conducted by employing content analysis. The questions generated by ChatGPT were analyzed by using content analysis to ensure that each question involved Higher Order of Thinking Skills (Cognitive Level 4, 5, and 6) as the prompt instructed. While statistical methods were used to analyze the quantitative data. The validity and reliability of the questions generated by ChatGPT were calculated by using SPSS version 29.

Population and Sample

The study targeted students enrolled in the English Education Study Program at the Faculty of Teacher Training and Education, Universitas Al Washliyah Medan, specifically those in the third, fifth and seventh semesters during the academic year 2023-2024. The population comprises 42 students, and through the utilization of random sampling technique involving a wheel spin, 25 students comprising of 10 semester 3 students, 10 semester 5 students, and 5 semester 5 students, were selected for the research. This sampling method aimed to ensure representation across different semesters, allowing for a diverse range of proficiency levels among the participants, a crucial aspect for the study’s objectives.

Research Instrument

To gather data, a set of reading comprehension test was administered through a paper-based-test, comprising multiple-choice questions with 4 options created by Chat-GPT 3.5. These questions were generated based on carefully selected academic texts, specifically chosen from the TOEFL ITP reading comprehension passages. The chosen academic passages consisted of three expository passages with diverse topics:

1. Passage 1 - Telescope Photography and Technology Impact:
   - ✓ Focus: Astronomy and telescope photography.
   - ✓ Main Topics: Direct photography in astronomy, use of glass plates, limitations of photography, technology impact on modern astronomy (radio and x-ray telescopes), and the role of image processing.
2. Passage 2 - Impressionism and Artistic Innovations:
   - ✓ Focus: Art and the Impressionist movement.
   - ✓ Main Topics: Emergence of Impressionism in 1874, dissatisfaction with the academic art establishment, technological innovations in art (new brushes, collapsible tin tubes, and a new palette of colors), and the impact of these innovations on the artistic style.
3. Passage 3 - Radiocarbon Dating and Tree Ring Dating:
   - ✓ Focus: Scientific methods for dating.
   - ✓ Main Topics: Radiocarbon dating and dendrochronology (tree ring dating) as tools for establishing a time spectrum, the process of tree ring formation, factors influencing ring thickness, and the correlation of growth rings between trees.

In summary, passage 1 is about astronomy and telescope technology, passage 2 is about the impressionist art movement and its technological influences, and passage 3 is about scientific dating methods, specifically radiocarbon and tree ring dating. The topics of the passages are different. While they all involved some aspect of technology or scientific methods, the specific subjects and contexts differ significantly.

For each passage, Chat-GPT 3.5 generated ten questions, resulting in a total of 30 test items. The questions aimed to comprehensively assess the sample's higher order of thinking skills covering Cognitive Levels 4-6, inspired by the revised version of Bloom Taxonomy. The prompt was carefully formulated in order to result optimal generated reading comprehension questions. The following is the prompt:

"Generate 10 multiple-choice questions based on the provided passage. Ensure that the questions involve higher-order thinking skills, specifically focusing on cognitive levels 4 to 6. For cognitive level 4 (analyzing), use operational verbs such as comparing, organizing, deconstructing, attributing, outlining, finding, structuring, and integrating. For cognitive level 5 (evaluating), use verbs like checking, hypothesizing, critiquing, experimenting, judging, testing, detecting, and monitoring. For cognitive level 6 (creating), employ verbs such as designing, constructing, planning, producing, inventing, devising, and making." The answer key along with the discussion for each question was also generated by using ChatGPT.

Techniques for Collecting Data

The students were administered reading comprehension tests comprising ChatGPT-generated multiple choice questions. The questions generated were qualitatively analyzed by employing content analysis. The students' responses were quantitatively analyzed to assess the validity and reliability of the test through statistical analyses. The test validity was assessed through Pearson Correlation formula to ensure internal validity. On the other

hand, the reliability of the test was determined using Cronbach's Alpha formula, focusing on internal consistency reliability.

The following presents more detailed description of the data collection technique of the mixed-method approach.

1. Qualitative Data Collection: Content Analysis

Procedure:

a. Text Selection: Choose three texts from TOEFL ITP reading comprehension test. The texts focus on diverse topics and are in the similar lengths. The texts covered different topics; Passage 1: Telescope Photography and Technology Impact, Passage 2: Impressionism and Artistic Innovations, and Passage 3: Radiocarbon Dating and Tree Ring Dating.
b. Question Generation: Utilize ChatGPT 3.5 to generate a set of 10 reading comprehension questions for each text totaling 30 multiple choice questions.
c. Content Analysis: Conduct a comprehensive examination of each question. Assess each of the question involved Higher order of Thinking Skills (HOTS) in line with the prompt instructed.
d. Coding Process: Develop a coding system to categorize questions based on predetermined criteria. There were three codes used representing each cognitive level: Cognitive Level 4 - Analysis (C4), Cognitive Level 5 - Sinthesize (C5), and Cognitive Level 6 - Create (C6).

2. Quantitative Data Collection: Statistical Analysis

Procedure:

a. Test Administration: Administer the 30 generated questions to the 25 participants through a paper-based-test.
b. Scoring: Score each correct response 1, and 0 for incorrect answer.
c. Validity Assessment: Utilize Pearson Correlation formula to examine the construct validity of the questions.
d. Reliability analysis: Employ Cronbach's Alpha formula to assess the internal consistency of the generated questions. Cronbach's Alpha is a measure of internal consistency, reflecting how well the items in a scale or test measure the same underlying construct
e. Comparison: Correlate qualitative findings with quantitative results to obtain a comprehensive understanding of the questions' effectiveness.

3. Integration

a. Synthesis: Synthesize qualitative and quantitative findings to provide a comprehension interpretation.
b. Recommendations: Draw conclusions based on the integrated results. Propose feasible recommendations for improving the reliability and validity of ChatGPT 3.5-generated reading comprehension questions for academic texts.

RESULTS AND DISCUSSION

Question Analysis

By conducting content analysis, the questions were examined to ensure that each one involved higher order of thinking skills within the Cognitive Levels 4-6, covering analysis, evaluation, and creation. Table 1 offers the results of analysis of each question based on the specified cognitive levels.

Table 1. Results of Analysis of Each Question Based on the Specified Cognitive Levels

Cognitive Level	Code	Questions	Number of Questions	Percentage (%)
Cognitive Level 4 - Analysis	C4	1,3,4,8,11,14,17,20,21,24,27,30	12	40
Cognitive Level 5 - Evaluating	C5	2,5,9,12,15,18,22,25,28	9	30
Cognitive Level 6 - Creating	C6	6,7,10,13,16,19,23,26,29	9	30

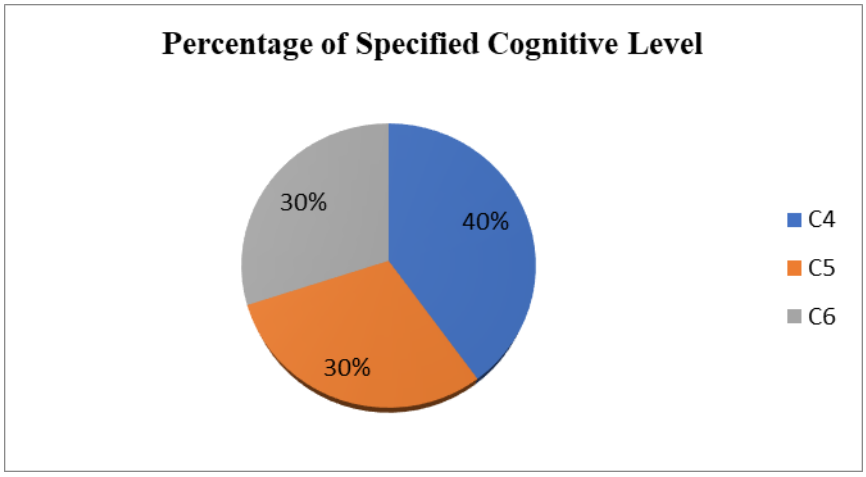

Figure 1. Number of Correct Answers per Question

Based on the analysis of the cognitive levels of the questions, it is evident that the questions are distributed across different levels of cognitive levels. The findings indicated that 40% of the questions were categorized under Cognitive Level 4 - Analysis, 30% under Cognitive Level 5 - Evaluating, and another 30% under Cognitive Level 6 - Creating. This distribution suggests a balanced representation of different cognitive levels within the assessment, which is essential for evaluating students' comprehensive understanding and application of knowledge.

The distribution of questions across these cognitive levels in accordance with the principles of Bloom's Taxonomy, which emphasizes the importance of assessing higher-order thinking skills such as analysis, evaluation, and creation alongside lower-order cognitive skills. This approach is crucial for promoting critical thinking and problem-solving abilities among students, as it encourages them to engage with the material at a deeper level and apply their knowledge in novel contexts.

Validity and Reliability Examinations

Validity Examination

The validity assessment, conducted through the Pearson Correlation formula, indicated that out of 30 questions generated by utilizing ChatGPT 3.5, 10 were found to be valid. According to Pearson Correlation standards, if the correlation coefficient is higher than r-table, the correlation is deemed statistically significant. This implies that the validity of the questions is established. Table 2 presents the results of the validity assessment.

Table 2. Validity Statistics

Question	r_{table}	$r_{counted}$	Interpretation
Question 1	0.381	0.164	Invalid
Question 2	0.381	0.289	Invalid
Question 3	0.381	0.331	Invalid
Question 4	0.381	0.374	Invalid
Question 5	0.381	0.503	Valid
Question 6	0.381	0.057	Invalid
Question 7	0.381	0.363	Invalid
Question 8	0.381	0.523	Valid
Question 9	0.381	0.533	Valid
Question 10	0.381	0.290	Invalid
Question 11	0.381	0.616	Valid
Question 12	0.381	0.339	Invalid
Question 13	0.381	0.227	Invalid
Question 14	0.381	0.204	Invalid
Question 15	0.381	0.260	Invalid
Question 16	0.381	0.005	Invalid
Question 17	0.381	0.484	Valid
Question 18	0.381	0.035	Invalid
Question 19	0.381	0.516	Valid
Question 20	0.381	0.315	Invalid
Question 21	0.381	0.380	Invalid
Question 22	0.381	0.599	Valid
Question 23	0.381	0.148	Invalid
Question 24	0.381	0.394	Valid
Question 25	0.381	0.351	Invalid
Question 26	0.381	0.654	Valid
Question 27	0.381	0.035	Invalid
Question 28	0.381	0.617	Valid
Question 29	0.381	0.766	Valid
Question 30	0.381	0.363	Invalid

The statistical analysis results indicate the validity of 10 questions, specifically those numbered 5, 8, 9, 11, 17, 19, 22, 26, 28, and 29. While, the remaining 20 questions did not meet the validity criteria.

Reliability Examination

The following tables are the outputs resulted from the reliability assessment by employing Cronbach's Alpha.

Table 3. Validity Statistics

Case Processing Summary			
		N	%
Cases	Valid	25	100.0
	Excluded[a]	0	.0
	Total	25	100.0

a. Listwise deletion based on all variables in the procedure.

Table 4. Validity Statistics

Reliability Statistics	
Cronbach's Alpha	N of Items
.671	30

The statistical analysis of reliability, as indicated by Cronbach's Alpha, resulted in a value of 0.671. Cronbach's Alpha is a measure of internal consistency, reflecting how well the items in a scale or test measure the same targeted construct. Generally, higher values of Cronbach's Alpha (closer to 1.0) indicate greater reliability. In this analysis, a value of 0.671 suggested a moderate level of internal consistency among the 30 items included in the analysis. While it was not exceptionally high, the value indicated a reasonable degree of reliability.

Research Findings

The following is the interpretation of the findings:

1. The results of content analysis suggested that the ChatGPT 3.5-generated reading comprehension questions have effectively met the prompt objective of constructing questions that involved higher order of thinking skills, specifically within cognitive Levels 4-6 of Blooms' Taxonomy. It was also found out that it effectively achieved a well-balanced cognitive level distribution with 40% at Cognitive Level 4 (Analysis), 30% at Cognitive Level 5 (Evaluation), and another 30% at Cognitive Level 6 (creation). The ChatGPT 3.5's ability to match with established educational frameworks indicated its effectiveness in generating questions that met recognized standards for cognitive complexity. Moreover, the ChatGPT 3.5's generated questions promoted critical thinking and problem-solving among students. This suggested that ChatGPT 3.5 was successful in encouraging students to think critically, analyze information, evaluate content, and create new ideas. It implied that the generated questions were designed to foster a deeper understanding of academic text. Thus, the interpretation revealed that ChatGPT 3.5 has effectively generated reading comprehension questions that meet the specified higher-order thinking skills (Levels 4-6).
2. The findings reveal a picture of ChatGPT 3.5's performance in generating reading comprehension questions. The valid questions, numbering 10 out of 30, demonstrate the ChatGPT 3.5's ability to generate content in line with the expected constructs. However, the majority of questions did not meet the validity criteria, suggesting a need for improvements in generating relevant and effective questions. The moderate reliability suggests a consistent internal structure among the questions, providing a basis for possible improvement. Overall, while ChatGPT 3.5 shows promise. Further adjustment and enhancement are crucial. to ensure a more reliable and valid tool for generating academic reading comprehension questions.

CONCLUSION

The ChatGPT 3.5-generated reading comprehension questions for academic texts align effectively with Cognitive Levels 4-6 of Bloom's Taxonomy. It effectively fulfills the research objective by generating reading comprehension questions that engage students in higher-order thinking skills within the specified cognitive levels. The results suggest its potential as a reliable tool for constructing assessments that assess and enhance students' cognitive abilities in academic contexts.

The study shows how ChatGPT 3.5 manages generating reading comprehension questions. Out of 30 questions, only 10 turned out valid, indicating that ChatGPT 3.5 can match with the intended goals. But most questions did not meet the validity criteria, suggesting a need for improvements in making relevant and impactful questions. The moderate reliability suggests a consistent structure among the questions, forming a basis for potential improvements. Overall, while ChatGPT 3.5 seems promising, there's a clear need for more adjustment and improvements. These adjustments are vital to make ChatGPT a more valid and reliable tool for consistently producing high-quality academic reading comprehension questions. The research findings provide insights that can guide the enhancement of AI models for better alignment with educational goals.

In light of the research finding, it is recommended to enhance the validity criteria for ChatGPT-3.5 generated questions by improving and expanding the assessment parameters.

REFERENCES

Cukurova, M., Kent, C., & Luckin, R. (2019). Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology*, *50*(6), 3032–3046.

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Medical Education*, *9*(1), e46885.

Hung, Y.-C., Chaker, S. C., Sigel, M., Saad, M., & Slater, E. D. (2023). Comparison of Patient Education Materials Generated by Chat Generative Pre-Trained Transformer Versus Experts: An Innovative Way to Increase Readability of Patient Education Materials. *Annals of Plastic Surgery*, *91*(4), 409–412.

Kim, Y.-S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology*, *57*(5), 718.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783.

Schiff, D. (2021). Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI & Society*, *36*(1), 331–348.

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, *100*(8), 3098–3101.

Vandiver, V. L. (2008). *Integrating health promotion and mental health: An introduction to policies, principles, and practices*. Oxford University Press.