Journal of English Language Teaching
Department of English Language Education
FBS Universitas Negeri Padang, E-ISSN: 2302-3198
https://jelt.ppj.unp.ac.id/index.php/jelt

An Analysis of ChatGPT-Generated Cloze Tests for Reading Assessment of College Students: Practicality, Validity, and Reliability

Muhammad Iqbal Julianda[1*], Dinovia Fannil Kher[1]
[1] Universitas Negeri Padang, Indonesia
*Correspondence Email: Iqbal.julianda65@gmail.com

Volume 15, Number 1, 2026, 27-34
DOI: https://doi.org/10.24036/s4x9fz12

Abstract:
The use of artificial intelligence in language assessment has increased interest in automated test generation. In reading assessment, cloze tests are widely used to measure comprehension through contextual processing. This study examines the practicality, validity, and reliability of a ChatGPT-generated cloze test for reading assessment of college students. Using a descriptive quantitative design, 25 students completed a 30-item cloze test generated by ChatGPT without human modification. Practicality was measured through a 20-item student perception questionnaire, while validity and reliability were analyzed using corrected item–total correlation and Cronbach's Alpha. The results indicate high practicality and acceptable reliability, although only a limited number of items were valid. These findings suggest that ChatGPT can support language test generation with proper statistical analysis.

Keywords:
ChatGPT, cloze test, practicality, validity, reliability, college students

1. INTRODUCTION AND LITERATURE REVIEW

Artificial intelligence (AI) has increasingly influenced various aspects of education, including instructional delivery, learning support, and assessment design. AI-powered systems are widely used to personalize learning, automate feedback, and assist teachers in developing instructional materials (Zawacki-Richter et al., 2019). In language education, the integration of AI has opened new possibilities for generating assessment instruments more efficiently than traditional manual methods. However, the rapid adoption of AI tools also raises questions regarding the quality and credibility of the assessments they produce.

One of the most prominent AI technologies in recent years is ChatGPT, a large language model developed by OpenAI that can generate coherent and contextually relevant text (Brown et al., 2020). In educational contexts, ChatGPT has been utilized as a learning assistant, a writing aid, and a tool for generating instructional materials and assessment items (Kasneci et al., 2023). Its ability to automatically generate test items, including cloze tests, offers potential benefits in reducing teachers' workload and accelerating test development. Nevertheless, the pedagogical soundness of such AI-generated assessments cannot be assumed without empirical evaluation.

Assessment quality is commonly evaluated through three fundamental principles: practicality, validity, and reliability (Bachman & Palmer, 1996; Brown, 2004). Practicality concerns the feasibility of administering a test in terms of time, clarity, and resources. Validity refers to the extent to which a test measures what it is intended to measure, while reliability addresses the consistency of test results. When assessment items are generated automatically by AI, these principles become even more critical, as the output is produced without direct pedagogical judgment.

Reading assessment aims to measure learners' ability to construct meaning from written texts by integrating linguistic knowledge, background knowledge, and contextual interpretation (Alderson, 2000). Reading comprehension involves multiple cognitive processes such as vocabulary recognition, syntactic parsing, and inferencing. Therefore, effective reading tests must require test-takers to process contextual information rather than rely on isolated language knowledge. One commonly used format in reading assessment is the cloze test, which measures readers' ability to use contextual cues to restore missing words in a passage.

In reading assessment, cloze tests have long been recognized as an effective tool for measuring comprehension by requiring test-takers to rely on contextual and semantic cues (Alderson, 1979). With the emergence of AI-based text generation, cloze tests can now be produced automatically. However, concerns remain regarding whether AI-generated cloze tests are aligned with learners' proficiency levels, display appropriate item functioning, and yield consistent measurement results (Yan et al., 2022).

Previous studies on ChatGPT in education have largely focused on its instructional benefits and ethical implications, often relying on qualitative analysis or expert judgment. Empirical studies examining the quality of ChatGPT-generated assessments using statistical analysis are still limited. Addressing this gap, the present study aims to examine the practicality, validity, and reliability of ChatGPT-generated cloze tests for reading assessment using a fully quantitative approach. By analyzing students' perceptions and test performance data, this study seeks to provide evidence-based insight into the extent to which ChatGPT-generated cloze tests can function as usable language assessment instruments.

2. METHOD

This study employed a descriptive quantitative design to examine the practicality, validity, and reliability of a ChatGPT-generated cloze test. The participants were 25 students from an English Language Education program who had completed an Intermediate Reading course. The study employed two research instruments. The first instrument was a 30-item ChatGPT-generated cloze test used to measure students' reading comprehension performance. The second instrument was a 20-item student perception questionnaire designed to measure the practicality of the test using a five-point Likert scale. While the cloze test data were analyzed for validity and reliability, the questionnaire responses were used exclusively to evaluate practicality. All data were analyzed using SPSS.

3. RESULTS AND DISCUSSION

3.1. Results

This section reports the results from two different data sources. The practicality results are based on the 20-item student perception questionnaire, whereas the validity and reliability results are based on students' responses to the 30-item cloze test.

1. Practicality

Table 1. Practicality Analysis

PRACTICALITY ANALYSIS

ITEM	MEAN	SD	CATEGORY
I1	3,56	1,356	HIGH
I2	2,96	1,457	MODERATE
I3	3,40	1,000	HIGH
I4	3,84	1,179	HIGH
I5	3,88	1,166	HIGH
I6	3,76	1,200	HIGH
I7	3,76	1,091	HIGH
I8	3,72	1,173	HIGH
I9	3,64	1,186	HIGH
I10	3,16	1,375	MODERATE
I11	3,48	1,194	HIGH
I12	3,52	1,327	HIGH
I13	3,52	1,262	HIGH
I14	3,68	1,145	HIGH
I15	3,72	1,275	HIGH
I16	3,60	1,118	HIGH
I17	3,64	1,221	HIGH
I18	3,72	1,173	HIGH
I19	3,76	1,165	HIGH
I20	3,88	1,236	HIGH

The analysis of the student perception questionnaire indicates that the ChatGPT-generated cloze test demonstrates a high level of practicality. Most respondents agreed that the test instructions were clear, the layout was easy to follow, and the allocated time was sufficient to complete the test. The overall mean score of the questionnaire falls within the high practicality category, suggesting that the test is feasible for classroom or research use without causing excessive burden to test-takers.

2. Validity

In language testing, validity can be categorized into content, face, criterion, and construct validity. The present study focuses on construct validity, which examines whether each test item functions consistently in measuring the same underlying construct. In this research, the construct refers to reading comprehension ability measured through a cloze test. Construct validity was analyzed using corrected item–total correlation, which indicates how well each item correlates with the overall test score. Items with correlation values equal to or higher than the critical r-value were considered valid.

Table 2. Validity Analysis

Validity **Interpretation**
[r ≥ 0.396 → valid (r table for N=25, α = 0.05)]

Question	Corrected Item-Total Correlation (r)	Status
Q1	.686	VALID
Q2	.245	INVALID
Q3	.579	VALID
Q4	.000	INVALID
Q5	.000	INVALID
Q6	.245	INVALID
Q7	.462	VALID
Q8	.598	VALID
Q9	.587	VALID
Q10	.648	VALID
Q11	.144	INVALID
Q12	.233	INVALID
Q13	.434	VALID
Q14	.096	INVALID
Q15	-.052	INVALID
Q16	.277	INVALID
Q17	.216	INVALID
Q18	.480	VALID
Q19	.073	INVALID
Q20	.284	INVALID
Q21	.398	VALID
Q22	-.179	INVALID
Q23	.285	INVALID

Q24	.261	INVALID
Q25	.345	INVALID
Q26	.215	INVALID
Q27	.324	INVALID
Q28	.185	INVALID
Q29	.303	INVALID
Q30	.107	INVALID

TOTAL: 9 VALID, 21 INVALIDS

Construct validity was examined using corrected item–total correlation. The results show that 9 out of 30 test items met the minimum validity criterion, while the remaining items did not demonstrate sufficient correlation with the total test score. This indicates that only a limited number of items functioned effectively in measuring the intended construct, whereas most items showed weak item discrimination. The dominance of invalid items suggests that the ChatGPT-generated cloze test, in its current form, requires substantial item revision before it can be considered an adequate reading assessment instrument.

3. Reliability

Table 3. Reliability Analysis

Reliability Statistics	
Cronbach's Alpha	N of Items
.771	30

Cronbach's Alpha Value	
$\alpha \geq 0.90$	Excellent
$0.80 \leq \alpha < 0.90$	Good
$0.70 \leq \alpha < 0.80$	Acceptable
$0.60 \leq \alpha < 0.70$	Questionable
$0.50 \leq \alpha < 0.60$	Poor
$\alpha < 0.50$	Unacceptable

Reliability analysis using Cronbach's Alpha indicates that the ChatGPT-generated cloze test achieved an acceptable level of internal consistency. The reliability coefficient suggests that the test items collectively measure a relatively consistent construct.

.

3.2. Discussion

The high level of practicality indicates that students perceived the ChatGPT-generated cloze test as clear, manageable, and appropriate for their reading level. This finding aligns with

previous studies reporting that AI-generated learning and assessment materials are generally well-received by students due to their clarity and efficiency (Kasneci et al., 2023). From an assessment perspective, this result extends prior research by demonstrating that ChatGPT can meet basic administrative and usability requirements for language testing, not only for instructional support.

In contrast, the construct validity results reveal a major limitation in the ChatGPT-generated cloze test, as only a small proportion of the items met the validity criterion. This indicates that ChatGPT did not consistently generate items aligned with the intended construct of reading comprehension. While the AI can produce grammatically correct and contextually appropriate texts, it does not automatically ensure that the items function well in measuring students' reading ability. This finding supports previous research suggesting that ChatGPT-generated assessment items may show uneven alignment with learning objectives when statistical considerations are not included in the generation process (Yan et al., 2022). Without systematic item analysis, such tests may include items that are too easy, too difficult, or not closely related to the intended skill. Therefore, the large proportion of invalid items highlights the need for post-generation statistical screening, suggesting that ChatGPT should be viewed as a supportive tool for initial item creation rather than a fully autonomous test developer, with human evaluation remaining essential to ensure test validity.

Despite the limited number of valid items, the reliability analysis indicates acceptable internal consistency. This finding suggests that the test items, when viewed collectively, functioned in a relatively coherent manner. Similar results have been reported in prior assessment research, where newly developed instruments demonstrated adequate reliability despite variability in item validity (George & Mallery, 2003). This result implies that reliability alone is insufficient as an indicator of assessment quality and must be interpreted alongside item-level validity analysis. Although the obtained reliability coefficient can be considered sufficient for classroom assessment and research purposes, higher reliability levels are generally required for high-stakes testing. Therefore, further refinement of the test items is necessary before the test can be used in formal or large-scale assessment contexts.

Overall, the findings of this study move the understanding of AI-generated language assessment forward by highlighting a critical distinction: ChatGPT-generated cloze tests may be practical and reliable yet still exhibit weaknesses in construct validity. This insight underscores the importance of empirical statistical evaluation when employing AI for test generation. Rather than replacing human involvement entirely, AI should be positioned as a supportive tool whose outputs require systematic analysis before being used for formal assessment purposes.

4. CONCLUSION

The This study investigated the practicality, validity, and reliability of a ChatGPT-generated cloze test for college students using a quantitative approach. The findings demonstrate that while the test is practically usable and shows acceptable internal consistency, its item-level validity remains limited. This indicates that the effectiveness of AI-generated assessments cannot be judged solely based on usability or overall reliability but must also consider how individual items function in measuring the intended construct.

The significance of this study lies in its empirical evaluation of AI-generated test instruments without human modification. By relying on statistical analysis rather than expert judgment, this research provides objective evidence of both the potential and limitations of ChatGPT in language test generation. The findings suggest that ChatGPT can serve as a supportive tool in developing assessment materials, particularly in reducing test construction

time, but should not be used as a standalone solution for formal language assessment without further evaluation.

In light of these findings, future research is recommended to explore strategies for improving the quality of AI-generated test items, such as prompt optimization, iterative item generation, or hybrid approaches that combine AI output with systematic item analysis. Further studies may also compare AI-generated tests with human-designed assessments across different language skills and proficiency levels to better understand the role of artificial intelligence in language assessment.

REFERENCES

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219–227. https://doi.org/10.2307/3586211

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference, 11.0 update* (4th ed.). Allyn & Bacon.

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 1*, Article 100001. https://doi.org/10.1016/j.caeai.2020.100001

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. Longman.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Yan, Z., Wang, T., & Wang, L. (2022). Rethinking the fairness of AI in education: Validity, bias, and transparency. British Journal of Educational Technology, 53(4), 838–856.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X. (2022). *ChatGPT user experience: Implications for education*. SSRN. https://doi.org/10.2139/ssrn.4312418