ROCZNIKI HUMANISTYCZNE
Tom LXXIII, zeszyt 11 – 2025

ZESZYT SPECJALNY / SPECIAL ISSUE
DOI: https://doi.org/10.18290/rh257311.8s

WOJCIECH MALEC

# A COMPARATIVE STUDY OF AI-GENERATED AND TEACHER-CONSTRUCTED LANGUAGE TESTS

**Abstract.** This article examines the quality of multiple-choice (MC) vocabulary tests generated with the help of artificial intelligence (AI) by comparing their psychometric properties to those of human-constructed tests. Two sets of criterion-referenced tests (CRTs), designed to assess vocabulary previously taught in class, were developed. In each set, one test was entirely generated by AI, while the other incorporated MC options either fully created by a human constructor or modified from the AI-generated options. The tests were administered to students of a high school. The analysis focussed on reliability estimates and item statistics, particularly those which are relevant to CRTs. The findings suggest that the use of AI significantly improved test practicality by reducing the time and effort needed to develop the tests, although human-constructed tests exhibited superior psychometric qualities.

**Keywords:** language assessment; criterion-referenced testing; multiple choice; automatic item generation; ChatGPT; test evaluation

## BADANIE PORÓWNAWCZE TESTÓW JĘZYKOWYCH GENEROWANYCH PRZEZ SZTUCZNĄ INTELIGENCJĘ I TESTÓW NAUCZYCIELSKICH

**Abstrakt.** Artykuł analizuje właściwości pomiarowe testów leksykalnych w formacie wyboru wielokrotnego (*multiple choice*, MC) generowanych przy użyciu sztucznej inteligencji (*artificial intelligence*, AI), poprzez porównanie ich wartości psychometrycznej z testami opracowanymi przez nauczyciela. Utworzono dwa zestawy klasowych testów sprawdzających (*criterion-referenced tests*, CRT), mających na celu ocenę znajomości słownictwa wcześniej omawianego na lekcjach. W każdym zestawie znalazł się jeden test w całości wygenerowany przez AI oraz drugi, w którym opcje odpowiedzi zostały przygotowane tradycyjnie lub zmodyfikowane na podstawie propozycji

WOJCIECH MALEC, PhD, Dr Litt., Assistant Professor at the John Paul II Catholic University of Lublin, Institute of Linguistics, Department of Theoretical Linguistics; correspondence address: Instytut Językowznawstwa KUL, Al. Racławickie 14, 20-950 Lublin, Poland; e-mail: wojciech.malec@kul.pl; ORCID: https://orcid.org/0000-0002-6944-8044.

I am grateful to the teachers of XXI Liceum Ogólnokształcące im. św. Stanisława Kostki in Lublin, Poland, who administered the vocabulary tests, and to the students who completed them.

AI. Testy przeprowadzono wśród uczniów szkoły średniej. Analiza skupiła się na współczynnikach rzetelności oraz wskaźnikach jakości zadań testowych, zwłaszcza tych, które są istotne w kontekście CRT. Wyniki wskazują, że wykorzystanie AI istotnie zwiększa praktyczność testowania, skracając czas i nakład pracy potrzebny do ich opracowania, jednak testy przygotowane przez nauczyciela odznaczały się lepszymi właściwościami psychometrycznymi.

**Słowa kluczowe:** ocenianie językowe; klasowe testy sprawdzające; zadania wyboru wielokrotnego; automatyczne generowanie zadań; ChatGPT; ewaluacja testów i sprawdzianów

## INTRODUCTION

The recent integration of artificial intelligence (AI) into learning and teaching has, by and large, had a beneficial impact on educational practices, particularly through the application of chatbots and generative AI (Jincheng et al., 2025). Much of this positive effect can be attributed to ChatGPT, a language generation model from OpenAI. It must be admitted that its use in education poses some challenges, for example because of concerns over malpractice and the risk of compromising academic integrity. On the other hand, ChatGPT has been reported to successfully supplement pedagogical activities by providing students with personalized learning or by supporting teachers in the development of instructional materials and assessment tasks; it can also communicate with the learners and provide feedback on their errors and the progress they make (Memarian & Doleck, 2023).

From a technical point of view, ChatGPT is a large language model (LLM) that can perform natural language processing (NLP) tasks. For example, it can interpret and generate text, hold natural and coherent conversations, write summaries, and translate languages. The ability to understand and produce human-like language is particularly useful in those educational contexts where efficient content generation is needed. For example, this AI model has been successfully used to generate learning objectives that were found to be measurable and relevant (Doyle et al., 2025), stories for reading comprehension that were found to be coherent, appropriate, and readable (Bezirhan & von Davier, 2023), as well as assessment items that properly adhered to the item writing guidelines (Zuckerman et al., 2023).

The application of AI to the writing of test items can be viewed as an instantiation of automatic item generation (AIG), which "is a technology-based innovation designed to scale the item development process so large numbers of high-quality items can be created efficiently and economically" (Sayin & Gierl, 2024, p. 5). It should be added that traditional AIG uses computer algorithms and strict psychometric principles to generate items on the basis of cognitive models and item templates (Gierl & Haladyna, 2013). On the other hand, ChatGPT-based AIG uses the generative capabilities of LLMs, which means that this system is probabilistic in nature and its outputs draw on patterns learned from large amounts of data. As a result, the items produced by ChatGPT-based AIG may not always meet strict psychometric standards, thus requiring human evaluation and modification. Despite these fundamental differences, within the broader AIG framework, the use of ChatGPT to generate test items can also be classified as AIG (e.g. Kıyak et al., 2024), or as AI-powered AIG (Shin & Lee, 2024).

Although AIG can dramatically enhance test practicality by reducing the resources required to construct test items, the quality of automatically generated items should not be taken for granted. However, evaluation of AIG items is, in fact, not consistently carried out, as reported by Circi et al. (2023). The easy availability of AI-based item generation systems further compounds the problem, particularly in classroom contexts, where test evaluation is relatively infrequent. This infrequency is due to the fact that many pre-service teachers do not receive any test development training, which leads to the omission of essential test construction procedures in actual classroom settings (Tewachew et al., 2024). As succinctly stated by Haynie (1992), “teachers lack sufficient training in test development [and] fail to analyze tests” (p. 26). Moreover, the ease of constructing tests with the aid of AI may even encourage the uncritical use of tests as ready-made products. In view of these concerns, there is a continuing need to conduct systematic investigations into the quality of classroom tests, especially those generated by AI.

## 1. EVALUATION OF AIG ITEMS

Evaluation of items generated by AIG has been conducted using various models. On the basis of two recent reviews of literature on automatic item generation in educational assessment (Circi et al., 2023; Song et al., 2025), it is possible to synthesize these models into three broad approaches. The first one may be termed *human-based evaluation*. It includes manual annotation by students, teachers, or experts using item rating questionnaires, blind review comparisons between AIG and human-constructed items, as well as the Turing test, in which humans attempt to detect AIG items. This approach is predominantly subjective, judgement-based, and qualitative. The second approach, *machine-assisted evaluation*, is principally objective, data-driven, and quantitative. It encompasses statistical item analyses, factor analyses of internal structure, and cosine similarity calculations. There are also hybrid man-machine models, in which human judgment is complemented by machine evaluation indicators. Finally, *effectiveness-based evaluation* focuses on measuring actual learning outcomes, and consists in determining whether using automatically generated tests results in significant knowledge gains. The approaches to evaluating AIG items may also be classified based on *who* evaluates the items (humans, machines, or both), *what* is evaluated (item quality in terms of content relevance, psychometric properties, or learning outcomes), and *how* the evaluation is conducted (subjectively or empirically).

A number of recent studies have applied these evaluation approaches, individually or in combination, to investigate the quality of AI-generated items in educational assessments. Some of these studies report that AI-powered AIG is on a par with human test development, or even superior. Other studies highlight the necessity of improving AI-generated items by human experts. There are also studies reporting mixed results.

One study, by Bhandari et al. (2024), compared questions generated by ChatGPT with traditional formative assessments and found no significant differences in difficulty and discrimination parameters between the two sets of items. It is noteworthy that ChatGPT items were even superior at differentiating the ability levels of the respondents. Moreover, AI-generated items demonstrated unidimensionality and, when used together with human-authored questions, did not affect the unidimensionality of the original item bank. Additionally, the learning objective distribution of ChatGPT-generated items showed greater similarity to the target lesson than to adjacent lessons.

Mendoza et al. (2024) evaluated items for a university entrance exam in written language using a detailed rubric. This study, which included items written by ChatGPT and humans, found that both sets had high rates of acceptance, although humans marginally outperformed ChatGPT. However, items produced by ChatGPT were, on the whole, more consistent and required fewer substantial edits. In general, the results of this study suggest that ChatGPT can generate test items of comparable quality to those written by human experts. A similar conclusion was reached by O (2024), who focussed on the parallelism between two test forms, one entirely produced by humans and the other created with partial assistance of AI. The results of statistical analyses confirmed the comparability of the two test forms.

Another study (Coşkun et al., 2025), in the context of medical education, relied on human-based evaluation and psychometric characteristics (item difficulty and point-biserial correlations) to investigate ChatGPT's ability to produce MC questions. The AI-generated items were positively evaluated by subject-matter experts. However, out of fifteen MC items, only six had acceptable point-biserial values, and just five items were deemed suitable for classroom tests. On the whole, the researchers acknowledged ChatGPT's potential for generating assessment questions.

A study by Shin and Lee (2023) on L2 assessment materials revealed that ChatGPT-generated reading passages and MC items were similar to human-created ones in terms of naturalness and flow. However, human-created items had more attractive MC options and higher completion levels—a term used to refer to the relevance of the items to the target passage as well as the clarity and content homogeneity of the options. While Shin and Lee appreciated the usefulness of AI-powered item generation, they also emphasized that "it is not yet possible to completely hand over L2 testing item creation to ChatGPT" (p. 35).

The feasibility of using ChatGPT to grade university exams has been explored by Flodén (2025). The results indicate that although AI-generated scores generally resembled those given by human graders, ChatGPT often awarded slightly higher marks. Moreover, complete agreement on grades was relatively low, with only 30% of exams receiving identical grades from both sources.

There are also studies reporting rather negative results. For example, the reliability of ChatGPT in perceiving and rating the complexity of writing prompts was found to be low compared to human raters in a study by Khademi (2023). Moreover, the test items generated by ChatGPT were frequently lengthy and showed similarities in meaning between the options (Aryadoust et al., 2024). Finally, Ngo et al. (2024) found that ChatGPT was able to generate only 19 correct items (with answers and explanations) out of a total of 60.

An important issue in the context of MC items is the quality of the distractors. Several studies have focussed on this problem. For example, Bitew et al. (2025) used qualitative and quantitative methods to evaluate ChatGPT's ability to generate MC distractors. According to the results, 53% of the generated distractors were of high quality. Another study (Malec, 2024) found that some of the distractors generated by ChatGPT for a multiple-choice vocabulary test were very ineffective, and follow-up queries failed to correct these errors in item construction. ChatGPT also produced some non-functional distractors, chosen by fewer than 5% of the test takers, in a study by Kıyak et al. (2024).

Finally, while a number of studies have compared AI-generated and human-constructed items, none of them seems to have specifically addressed criterion-referenced assessments. Given that criterion-referenced testing (CRT) is generally more appropriate in the language classroom than norm-referenced testing (NRT) (e.g. Brown & Hudson, 2002), it is important to fill this research gap by examining the quality of items designed for CRT interpretations.

This brief review shows that further research is needed to assess the quality of AI-generated items. Despite the promise of AI-based solutions, traditional concerns related to the appropriateness and usefulness of test scores continue to persist.

## 2. THE STUDY

Building on a previous study (Malec, 2024) that focussed on MC distractors, the present research uses part of the same dataset, supplemented with additional data, to explore the quality of entire items and tests produced with the aid of AI by comparing them to those developed by a human test constructor. The assessment instruments were designed for classroom use and intended for criterion-referenced score interpretations. The study was divided into two parts, each comprising two vocabulary tests.

### 2.1 METHOD

The participants in this study were high school students, aged 16–18, from Lublin, Poland, learning English at an advanced level corresponding to B2+/C1 on the CEFR scale. The group that participated in the first part of the study consisted of 142 students, whereas the group in the second part was divided into two subgroups (37 and 39 students, respectively). The students were not randomly assigned to the study groups, which could be seen as a limitation. For logistical reasons, they completed the tests during their regular instructional hours.

The instruments developed for the purpose of this study were MC vocabulary tests designed to assess students' knowledge of lexical items previously taught in class. In Part 1, the tests targeted vocabulary drawn from a single lesson in *Speakout Advanced*, taught immediately before the test administration. In Part 2, the tests covered vocabulary from the entire *New Password*

*B2+/C1* coursebook, which the students had completed as part of their school curriculum. The tests in Part 1 each consisted of 15 items, while those in Part 2 included 20 items. Every MC item consisted of a context sentence (stem) intended to reflect the meaning and use of the target lexical item, followed by three answer options—one correct answer (the key) and two distractors. The three-option format was chosen in line with frequent recommendations in the literature (e.g. Rodriguez, 2005). The cut-off score for each test was set at 50%, which the teachers considered an appropriate threshold for mastery at the students' proficiency level.

The following labels are used for the four tests:

– T1-AI: Test 1A (AI-generated),
– T1-H: Test 1B (human-created),
– T2-AI: Test 2A (AI-generated),
– T2-HM: Test 2B (human-modified AI test).

All tests were constructed in several stages. First, 30 lexical items were selected for Part 1 (15 for T1-AI and 15 for T1-H) and 20 for Part 2 (shared between T2-AI and T2-HM). Next, one context sentence was generated for each word or phrase using an AI-powered online platform, Twee (twee.com, accessed February 14, 2024). Then, ChatGPT (chat.openai.com, accessed February 14, 2024) was prompted to generate distractors for each item in T1-AI and T2-AI. The prompt contained detailed instructions on how to produce plausible yet plainly incorrect distractors (see Malec, 2024, for details). T1-H included traditional, human-written distractors, while T2-HM had distractors that were partly modified based on the original AI-produced ones. These modifications were made by the teacher-researcher conducting the study, and were guided by human judgement and classroom testing experience. Finally, the tests were assembled on WebClass (webclass.co) and administered online to the participants.

Following test administration, WebClass was also used to conduct most of the statistical analyses (see also Malec, 2025). Although test and item statistics may not be the most important consideration in the context of CTR, they still provide useful additional information regarding the quality of individual items and entire tests (Brown & Hudson, 2002; Bachman, 2004; Brown, 2014). For individual items, the *B*-index was used as a CTR equivalent of NRT discrimination indices. This statistic indicates how effectively a given item distinguishes between masters (students who passed the test) and non-masters (students who failed). For completeness and comparison, however, the point-biserial correlation coefficient was also included in the analysis. At the test level, the most appropriate CRT reliability (or, rather, dependability) estimate is the phi coefficient (Φ). To show the differences between CRT and NRT contexts, Cronbach's alpha (α) is also reported. Moreover, the consistency of mastery/non-mastery classifications was assessed using two indices, phi lambda ($\Phi_\lambda$) and kappa squared $\kappa^2$, both of which indicate the accuracy and consistency of student classifications relative to the established cut scores. In addition, the equivalence of test forms in each part of the study was examined using *t*-tests and the Two One-Sided Tests (TOST) procedure, originally proposed by Schuirmann (1987). The equivalence bounds were 0.75 for the first test pair and 1.00 for the second, in both cases corresponding to a 5% difference in total scores. This difference was assumed to be acceptable given that each grade band represented a 10% range of the total score. In other words, a 5% change in test scores would not, on average, result in a change of score-based decisions.

### 2.2 Results and discussion

The results are presented in two sections: test-level analysis and item-level analysis. The former focusses on general descriptive statistics, test form equivalence, and reliability/dependability estimations, whereas the latter is primarily concerned with item facility and discrimination.

#### 2.2.1 Test-level analysis

Descriptive statistics and reliability/dependability estimates for the four test forms are given in Table 1.

Table 1. Descriptive statistics and reliability/dependability estimates

| Statistic | T1-AI | T1-H | T2-AI | T2-HM |
|---|---|---|---|---|
| Number of test takers (*n*) | 142 | 142 | 37 | 39 |
| Number of items (*k*) | 15 | 15 | 20 | 20 |
| Cut score [test] ($\lambda_t$) | 7.5 | 7.5 | 10 | 10 |
| Cut score [item] (λ) | 0.5 | 0.5 | 0.5 | 0.5 |
| Mean ($\bar{x}$) | 7.53 | 8.94 | 13.46 | 16.56 |
| Mean of proportion scores ($\bar{x}_p$) | 0.50 | 0.60 | 0.67 | 0.83 |
| Standard deviation (*SD*) | 2.23 | 2.56 | 3.04 | 3.39 |
| Min. | 2 | 1 | 6 | 6 |

| | | | | |
|---|---|---|---|---|
| Median | 8 | 9 | 14 | 18 |
| Max. | 12 | 14 | 19 | 20 |
| Skewness | −0.588 | −0.930 | −0.459 | −1.327 |
| Kurtosis | −0.073 | 1.172 | −0.263 | 1.419 |
| Cronbach's alpha (α) | .488 | .567 | .654 | .796 |
| Standard error of measurement (*SEM*) | 10.64% | 11.24% | 8.95% | 7.64% |
| Phi coefficient (Φ) | .398 | .522 | .593 | .791 |
| Phi lambda ($\Phi_\lambda$) | .258 | .604 | .802 | .956 |
| Kappa squared ($\kappa^2$) | .488 | .672 | .851 | .958 |
| Standard error for absolute decisions ($SEM_{abs}$) | 12.76% | 12.30% | 10.20% | 7.76% |

Based on the descriptive statistics, the scores on T1-H and T2-HM were higher compared to T1-AI and T2-AI, respectively. Furthermore, both "human" versions of the tests had more dispersed score distributions, were more negatively skewed, and had higher kurtosis, which indicated a greater frequency of high scores. T2-HM, in particular, had an elevated median, which suggested a ceiling effect, with many students achieving near-perfect scores.

A paired-samples *t*-test comparing T1-AI and T1-H showed a statistically significant difference in scores, $t(141) = -6.72$, $p < .001$, which indicated a lack of score equivalence between the two forms. Similarly, for T2-AI and T2-HM, an independent-samples *t*-test with Welch's correction (due to unequal variances) also revealed a significant difference, $t(73.79) = -4.21$, $p < .001$. These results were confirmed by the TOST analysis, which revealed that the differences in scores were larger than the predefined equivalence bounds, thus indicating a lack of score equivalence between the test forms in both pairs. A summary of TOST results is given in Table 2.

Table 2. Summary of TOST results

| Comparison | Equivalence bounds | Mean difference | 90% CI of difference | TOST result |
|---|---|---|---|---|
| T1-AI vs T1-H (paired) | ±0.75 | −1.415 | [−1.76, −1.07] | Not equivalent |
| T2-AI vs T2-HM (independent) | ±1.00 | −3.105 | [−4.34, −1.88] | Not equivalent |

The reliability of the tests for NRT interpretations, represented by Cronbach's alpha (α), and the dependability of domain score estimates for

CRT interpretations, represented by the phi coefficient (Φ), were both higher for the second test in each pair. This means that the AI-generated forms exhibited greater measurement error. A similar pattern was observed in the case of decision consistency, as estimated by phi lambda ($\Phi_\lambda$) and kappa squared ($\kappa^2$). Both indices increased substantially from the AI-generated to the human-developed versions, which suggested improved agreement between observed and true classifications with respect to the cut score.

### 2.2.2 Item analysis

While NRT items should ideally be of intermediate difficulty, since such items typically have better discrimination (Brown, 2012), item facility (*IF*) values can also provide useful information in CRT contexts. In particular, items which are at the extreme ends of the facility scale often require careful qualitative analysis. In this study, three items (all AI-generated) had *IF* values below .20. One example is given below:

(1) *I saw him ___ to me from across the room.*
A. waving
B. beckoning
C. shouting

The extremely low facility index (*IF* = .05) for the item in (1) was likely due to the inclusion of inappropriate distractors. While *beckoning* was intended as the correct answer, both *waving* and *shouting* may have appeared lexically acceptable to the test takers, making the item unnecessarily difficult and potentially misleading.

At the other end of the facility scale, several items had very high *IF* values. Two items with the highest facility (*IF* = .97) were found in T2_AI, and one such item appeared in T2_HM. One of these items is shown below:

(2) *It's my birthday next week, so I want to ___ a party.*
A. give
B. catch
C. throw

The AI-generated item in (2) also lacked effective distractors. Specifically, while *throw* was the intended correct answer, *give* was not clearly incorrect in the context given. This probably contributed to confusion, as reflected in the item's negative CRT discrimination (*B* = –.03), which indicates that Option A was selected by a test taker who passed the test. In the human-modified version of this test, the distractors were changed to *deliver* and *issue*. Interestingly, this revision did not affect the item's facility (*IF* remained at .97), but it led to a marked improvement in discrimination, with the *B*-index rising to .50.

Item discrimination analysis included the calculation of the point-biserial correlation coefficient (*PB*) for NRT interpretations and the *B*-index for CTR interpretations. A number of threshold values for interpreting these indices have been proposed in the educational measurement literature, with .30 frequently cited as a benchmark for acceptable discrimination (e.g. Ebel, 1954; Bachman, 2004). A summary of item discrimination analysis is presented in Table 3.

Table 3. Summary of item discrimination analysis

| Statistic | T1-AI | T1-H | T2-AI | T2-HM |
|---|---|---|---|---|
| Mean *PB* | .33 | .39 | .36 | .46 |
| Min. *PB* | −.11 | .16 | .00 | .18 |
| Max. *PB* | .56 | .57 | .57 | .70 |
| % *PB* ≥ .30 | 67% | 73% | 60% | 90% |
| Mean *B* | .24 | .32 | .31 | .48 |
| Min. *B* | .00 | .15 | −.24 | .26 |
| Max. *B* | .48 | .47 | .76 | .95 |
| % *B* ≥ .30 | 47% | 73% | 55% | 80% |

As can be seen in Table 3, the human-constructed (T1-H) and human-modified (T2-HM) items outperformed their AI-generated counterparts. Among the four tests, T2-HM demonstrated the strongest item discrimination overall, with the highest means of *PB* and *B*. This test also had the largest proportion of items that were equal to or above the .30 threshold. In contrast, T1-AI showed the weakest performance, as it had only 67% items with *PB* ≥ .30 and 47% items with *B* ≥ .30. These results suggest that human involvement generally led to items that were more effective at distinguishing between students based on ability and mastery.

## 3. LIMITATIONS

This study has some notable limitations. First, the number of students who completed the tests in Part 2 was relatively small. While such sample sizes are common in classroom-based research, and thus reflect real instructional contexts, they may limit the reliability and generalizability of the test and item statistics. In Part 1, however, the sample size was more adequate and provides greater confidence in the results.

Second, the study was conducted at the beginning of 2024, when GPT-3.5 was freely available. Since then, however, newer models have been introduced. These are likely to perform better in generating test items. According to ChatGPT's own comparison, GPT-4-turbo offers improved semantic understanding, greater sensitivity to subtle lexical distinctions, and better contextual awareness. These enhancements could result in distractors that are more plausible and challenging, potentially improving the overall quality of AI-generated test items.

Nevertheless, when prompted to generate distractors for one of the items used in this study, ChatGPT-4 suggested a distractor that was a close synonym of the correct answer (see Table 4). In the context given, this distractor could not be considered plainly incorrect, and its inclusion in the test would violate one of the fundamental multiple-choice item-writing guidelines. In view of this, it appears that, despite being more advanced than GPT-3.5, GPT-4 may still generate inappropriate distractors. Although the model was able to correct itself, this example emphasizes the importance of human oversight in the item development process.

Table 4. Part of a conversation with ChatGPT-4

| | |
|---|---|
| **User:** | What distractors do you suggest for the word in square brackets in the following sentence:<br>*She was [banking] on him to help her succeed in the business venture.* |
| **ChatGPT:** | For *banking* in this context (meaning "relying" or "depending on"), the distractors should be words that seem plausible but do not fit the intended meaning. Here are some good options:<br>Distractors:<br>1. *gambling* – Similar in meaning but implies risk and uncertainty rather than trust.<br>2. *counting* – Close in meaning but often used in a different structure (*counting on* vs. *banking on*).<br>3. *saving* – Related to *banking* in the financial sense, creating a semantic trap.<br>4. *investing* – Fits the business context but suggests putting in resources rather than relying on someone. |

| | |
|---|---|
| **User:** | Isn't *counting on* the same as *banking on*? |
| **ChatGPT:** | You're absolutely right – *counting on* is a direct synonym of *banking on*, so it wouldn't work as a distractor. My mistake!<br>Here's a revised list of better distractors for *banking on*:<br>1. *hoping* – Seems similar but lacks the sense of expectation or reliance.<br>2. *guessing* – Implies uncertainty rather than trust.<br>3. *waiting* – Suggests passivity rather than dependence.<br>4. *saving* – Plays on the financial meaning of *banking* but doesn't fit the context. |

## CONCLUSION

This article has addressed a vital concern in education, namely the use of artificial intelligence to generate test items. While a number of previous studies have already investigated this issue from various perspectives, none of them has specifically examined the quality of AI-generated items for criterion-referenced testing. The present study focused on classroom-based vocabulary assessment and compared AI-generated items and tests to those constructed or revised by humans using selected item- and test-level statistics. In addition to quantitative analyses, a qualitative assessment of items was conducted.

The findings indicate that the ChatGPT-generated tests were clearly outperformed by human-developed tests in terms of NRT reliability, CRT dependability, decision consistency, as well as item discrimination. In this particular context, the core of the problem with AI-produced items was the quality of multiple-choice distractors (see also Malec, 2024). In attempting to suggest plausible but incorrect options, ChatGPT occasionally generated distractors that were not plainly wrong, thereby misleading the test takers and compromising item discrimination.

On the other hand, it would be unfair not to acknowledge the potential of AI as a support tool for test developers, particularly at the stage of item writing, which includes the construction of distractors. As long as the automatically generated items are subjected to careful human review, the quality of tests and items, and thus the appropriateness of score interpretations, should not be jeopardized.

In light of the findings of this study, future research should focus on ways to refine AI systems to make them suitable for educational assessment. It is also important to develop guidelines for making the best possible use of AI-generated content in test development.

## REFERENCES


Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, *6*, Article 100204. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100204

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, Article 100161. https://doi.org/https://doi.org/10.1016/j.caeai.2023.100161

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence*, *7*, Article 100284. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100284

Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2025). Distractor generation for multiple-choice questions with predictive prompting and large language models. In R. Meo & F. Silvestri (Eds.), *Machine learning and principles and practice of knowledge discovery in databases* (pp. 48–63). Springer Nature Switzerland.

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 323–335). Routledge.

Brown, J. D. (2014). Score dependability and decision consistency. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, Chap. 71, pp. 1182–1206). John Wiley & Sons.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*, Article 858273. https://doi.org/10.3389/feduc.2023.858273

Coşkun, Ö., Kıyak, Y. S., & Budakoğlu, I. İ. (2025). ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher*, *47*(2), 268–274.

Doyle, A., Sridhar, P., Agarwal, A., Savelka, J., & Sakr, M. (2025). A comparative study of AI-generated and human-crafted learning objectives in computing education. *Journal of Computer Assisted Learning*, *41*(1), Article e13092. https://doi.org/https://doi.org/10.1111/jcal.13092

Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, *14*, 352–364.

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201–224. https://doi.org/https://doi.org/10.1002/berj.4069

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Haynie, W. J. (1992). Post hoc analysis of test items written by technology education teachers. *Journal of Technology Education*, *4*(1), 26–38.

Jincheng, Z., Thada, J., & Rukthin, L. (2025). Meta-analysis of artificial intelligence in education. *Higher Education Studies*, *15*(2), 189–210.

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching*, *6*(1), 75–80.

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*. https://doi.org/10.1007/s00228-024-03649-x

Malec, W. (2024). Investigating the quality of AI-generated distractors for a multiple-choice vocabulary test. In O. Poquet, A. Ortega-Arranz, O. Viberg, I.-A. Chounta, B. McLaren, & J. Jovanovic (Eds.), *CSEDU 2024: Proceedings of the 16th International Conference on Computer Supported Education – Volume 1* (pp. 836–843). SCITEPRESS.

Malec, W. (2025). Validating classroom tests on WebClass. In M. Bloch-Trojnar, A. Bloch-Rozmej, & E. Cyran (Eds.), *Form, function, and learning: Linguistic studies in honour of Professor Anna Malicka-Kleparska from her students, colleagues, and friends* (pp. 205–220). Wydawnictwo Werset.

Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans*, *1*(2), Article 100022. https://doi.org/https://doi.org/10.1016/j.chbah.2023.100022

Mendoza, K. K. R., Zúñiga, L. H. P., & López García, A. Y. (2024). Creación y jueceo de ítems: ChatGPT como diseñador y juez [Item creation and judging: ChatGPT as designer and judge]. *Texto Livre*, *17*, Article e51222. https://doi.org/10.1590/1983-3652.2024.51222

Ngo, A., Gupta, S., Perrine, O., Reddy, R., Ershadi, S., & Remick, D. (2024). ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology*, *11*(1), Article 100099. https://doi.org/https://doi.org/10.1016/j.acpath.2023.100099

O, K.-M. (2024). A comparative study of AI-human-made and human-made test forms for a university TESOL theory course. *Language Testing in Asia*, *14*(1), Article 19. https://doi.org/10.1186/s40468-024-00291-3

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, *43*(1), 5–18. https://doi.org/https://doi.org/10.1111/emip.12590

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, *27*(3), 27–40. https://doi.org/10.1186/s40468-024-00291-3

Shin, D., & Lee, J. H. (2024). AI-powered automated item generation for language testing. *ELT Journal*, *78*(4), 446–452. https://doi.org/10.1093/elt/ccae016

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. *Interactive Learning Environments*, 1–20. https://doi.org/10.1080/10494820.2025.2482588

Tewachew, A., Shiferie, K., & Tefera, E. (2024). Practices of EFL teachers in test construction. *Cogent Education*, *11*(1), Article 2412496. https://doi.org/10.1080/2331186X.2024.2412496

Zuckerman, M., Flood, R., Tan, R. J. B., Kelp, N., Ecker, D. J., Menke, J., & Lockspeiser, T. (2023). ChatGPT for assessment writing. *Medical Teacher*, *45*(11), 1224–1227. https://doi.org/10.1080/0142159X.2023.2249239