O *Language Testing in Asia* (2024) 14:19
https://doi.org/10.1186/s40468-024-00291-3

Language Testing in Asia

RESEARCH Open Access

A comparative study of AI-human-made and human-made test forms for a university TESOL theory course

Kyung-Mi O[1*]

*Correspondence:
kmo@dongduk.ac.kr

[1] Dongduk Women's University, 60 Hwarang-ro 13-gil, Seongbuk-gu, Seoul 02748, South Korea

Abstract

This study examines the efficacy of artificial intelligence (AI) in creating parallel test items compared to human-made ones. Two test forms were developed: one consisting of 20 existing human-made items and another with 20 new items generated with ChatGPT assistance. Expert reviews confirmed the content parallelism of the two test forms. Forty-three university students then completed the 40 test items presented randomly from both forms on a final test. Statistical analyses of student performance indicated comparability between the AI-human-made and human-made test forms. Despite limitations such as sample size and reliance on classical test theory (CTT), the findings suggest ChatGPT's potential to assist teachers in test item creation, reducing workload and saving time. These results highlight ChatGPT's value in educational assessment and emphasize the need for further research and development in this area.

Keywords: AI-assisted test development, AI test item generation, AI-assisted assessment, AI-human collaboration in tests, ChatGPT in assessment, Classical test theory, Parallel test forms, AI-human-made vs. human-made

Introduction

ChatGPT reached approximately 180.5 million users 5 days after its launch in November 2022 (Shewale, 2024). This achievement starkly contrasts with Instagram, which took two and a half months to reach a similar number, and Netflix, which took three and a half years to accumulate 1 million users. This rapid growth is partly attributed to its natural language processing (NLP) capabilities, simplifying user interactions by eliminating the need for complex programming language. Users receive responses in natural language in a conversational form akin to interacting with other individuals.

Moreover, ChatGPT's versatility has aroused the interest of professionals across various fields, prompting them to explore its potential applications within their respective domains. Its ability to engage users on diverse topics and address their varied questions and needs has made it particularly appealing. ChatGPT has been utilized in educational contexts, including in the creation of assessment items to meet the demands of continuous testing and frequent evaluations.

In educational settings, there has been a growing demand for the rapid generation of assessment items to accommodate continuous testing requirements (Kurdi et al., 2019). This shift has posed challenges to traditional test item creation methods and to the maintenance of test item bank stability (Circi et al., 2023). Finding high-quality test items has consistently proven difficult, with the manual creation of items being time-consuming and costly (Gehringer, 2004).

Similarly, the creation of teacher-made tests for classroom use faces limitations due to restricted financial and human resources. Many teachers encounter difficulties designing quality items for each assessment, often resorting to item reuse across terms (Gehringer, 2004; Wellberg, 2023). However, this practice may lead to issues such as students memorizing answers without engaging with the content and the risk of cheating through item over-exposure (Circi et al., 2023; Gehringer, 2004).

Given the significant implications of classroom assessments on students, particularly in secondary education, where they can influence postsecondary pathways, the need for robust and diverse item banks becomes apparent. Recognizing this need for item development, employing technological assistance to alleviate the burden on teachers appears to be a logical progression (Aryadoust et al., 2024; Settles et al., 2020; Swiecki et al., 2022). Inspired by this necessity, this study investigates whether ChatGPT 3.5 can help teachers design tests. Specifically, the study examines whether artificial intelligence (AI) assisted test forms can mirror human-made test forms. Consequently, two research questions have been formulated for the study:

Are the contents of each test item in the AI-human-made and human-made test forms parallel?

Are the AI-human-made and human-made test forms parallel?

Literature review

ChatGPT and chatbots in the field of language learning and assessment

ChatGPT, one of the most advanced and capable AI chatbots, was developed for users' conversational interactions by OpenAI, a nonprofit organization founded in 2015. OpenAI's development of ChatGPT aimed to create an advanced artificial intelligence chatbot capable of natural and coherent conversation, serving various purposes such as assisting users, providing information, and facilitating communication. As the most recent advancement within the systems category called "chatbots" (Taecharungroj, 2023), the essence of AI chatbots is rooted in employing natural language processing models, which enable computers to comprehend human language (Nagarhalli et al., 2020). Due to its ability to engage in natural and coherent conversations with users and its resourcefulness in diverse areas, many users have employed and studied its application across various sectors. ChatGPT has made its presence known to learners and instructors, and the advantages of using AI-based chatbots in the field of language learning and testing have been extensively explored.

Firstly, the chatbot aids in assessing and practicing aural and oral language skills. According to Aryadoust et al. (2024), creating reliable and beneficial listening assessments is often costly, cumbersome, and labor-intensive, leading many teachers to adopt commercialized tests. Moreover, selecting appropriate listening materials for students can pose challenges. In such cases, teachers can utilize AI-based tools to customize

natural conversational test tasks with multimodal materials for their students (Jeon et al., 2023). Although ChatGPT does not directly process spoken input, learners can utilize external speech recognition services to transcribe spoken language into text. ChatGPT then leverages this text input to provide feedback and responses. Furthermore, learners can potentially benefit from ChatGPT's capability to engage in text-based exchanges resembling written conversations. Although ChatGPT currently lacks the ability to replicate the fluidity and nuances of natural spoken language, it can still provide learners with opportunities for practice sessions and prompt feedback, albeit in a written format. This functionality serves as a viable alternative to tutoring (Hong, 2023) and promotes autonomous learning experiences (Jeon et al., 2023).

Secondly, AI-powered tools, including ChatGPT, significantly impact essay scoring and feedback. Chatbots can be preprogrammed with consistent and objective models (Kooli, 2023; Pranav, 2022), enabling the assessment of test takers' responses based on predetermined rubrics (Abida et al., 2023). While ChatGPT may occasionally misinterpret information and inadvertently introduce biases, which may go unnoticed, with responsible usage, users may mitigate such errors and biases (Meyer, et al. 2023). Moreover, chatbots can analyze the content and structure of learners' responses and identify errors within a short timeframe. This instantaneous feedback enables instructors to support learners in enhancing the quality of their writing and learning outcomes (Benali, 2021) while alleviating the burden of excessive grading (Abida et al., 2023; Thao, 2023).

Lastly, AI-powered tools can automate the generation of test items and learning assessments. ChatGPT can automate assessments, including continuous feedback-integrated assessment (Rudolph et al., 2023) and the generation of test items with original reading passages (Shin & Lee, 2023) or listening scripts (Aryadoust et al., 2024). The AI-powered item generation has attracted considerable interest among researchers in the field of language learning and assessment.

AI item generation

Although the terms "AI Item Generation" and "Automatic Item Generation" (AIG) may sometimes be used interchangeably because both approaches employ computer algorithms for item generation, they can be slightly different. Item generation using AI typically employs advanced natural language processing techniques for generating text-based items that do not rely on psychometric models (Gierl et al., 2012) as AIG does. AIG focuses on psychometric principles for generating test items, suggesting that, in automatic item generation, constructing test items adheres to established cognitive models, representing the knowledge and skills necessary to be measured. Test developers analyze and break down the logical reasoning process when formulating cognitive models (Pugh et al., 2016). This data is then put into a computer program with algorithms to produce test items (Pugh et al., 2016). This method generates test items using a consistent stem, typically in multiple-choice questions, with variations based on predetermined topics (Gierl et al., 2012).

Many items can be generated for a specific topic based on a single cognitive model (Gierl et al., 2012), and the models are standards in measurement theories, allowing the developed tests to serve the assessment purposes of validity, reliability, fairness, and quality. Moreover, AIG is known to make test and assessment development easier by

making it quicker to create items, reducing the cost of item creation, helping to continuously and rapidly develop a large pool of items, and tailoring items to fit individual learning needs for better outcomes (Circi et al. 2023).

Despite the advantageous features of AIG, the practical application of AIG in classroom settings poses challenges for educators. The complexity lies in teachers needing to conceptualize the test model, deconstruct assessment domains, scrutinize test specifications, and translate natural language descriptions into computer algorithms. In contrast, AI item generation using ChatGPT has many strengths. With the aid of natural language processing (NLP) techniques, educators can readily generate test materials by inputting prompts directly into the text input field provided on the platform. A fundamental objective of AIG, which is to efficiently offer test takers unique yet conceptually aligned assessments (e.g., Pugh et al., 2016), appears to be achievable for classroom teachers through ChatGPT.

Furthermore, ChatGPT can generate coherent and contextually appropriate prompt responses (Brown et al., 2020). These responses can demonstrate lexical and syntactic sophistication (Aryadoust et al., 2024), enabling test developers to customize items according to the proficiency levels of diverse test takers (Vajjala & Meurers, 2012). However, since ChatGPT is not specifically tailored for psychometric purposes, additional considerations and adaptations may be necessary to ensure the suitability and validity of generated test items for assessment purposes.

Test item generation and ChatGPT

Due to the recent introduction of ChatGPT, there has been limited research so far. However, researchers have increasingly become interested in exploring its potential applications, some of which will be discussed in this paper.

Aryadoust et al. (2024) investigated the potential of ChatGPT 4 in developing listening assessments to resolve the complexity and high costs of creating tests for individuals with varying proficiency levels. Prompt engineering and fine-tuning techniques were employed to create listening scripts and test items catering to various proficiency levels, encompassing academic, low, intermediate, and advanced levels, with 24 topics selected from academic listening tests to ensure consistency in the study. Two analyses were performed to analyze the output quality: one focused on the words in the scripts using CohMetrix and Text Inspector, and the other examined how different the topics were and if there was any overlap in the test questions. The findings suggest that although ChatGPT 4 consistently developed scripts with noticeable differences in wording, the resulting test questions were frequently lengthy and showed similarities in meaning between choices, which were affected by the topic. The study demonstrates the current stage of ChatGPT in test generation, indicating that although the application can reduce test development expenses, it still necessitates human supervision and expertise in refining prompts.

Kiyak, Coşkun, Budakoğlu, and Uluoğlu (2024) explored utilizing ChatGPT for generating case-based multiple-choice questions in medical studies. For the study, the researchers generated 10 multiple-choice questions on hypertension. Two of the 10 were selected by an expert panel and used without revision on a medical school exam administered to 99 medical students. Based on the data gathered, the researchers reviewed the psychometric characteristics based on classical test theory (CTT), including item

difficulty, item discrimination, and functionality of the options. The two items exhibited acceptable levels of item discrimination, suggesting the potential of ChatGPT in test development. While the study has limitations due to its narrow scope of analyzing only two test items and including some non-functional options, it suggests that ChatGPT could facilitate test-making. This work was immediately followed by Kiyak and Kononowicz (2024), who developed a customized version of GPT called the Case-based Multiple-Choice Question (MCQ) Generator to serve the practical needs in medical education for saving time and managing ChatGPT's exposure to limited medical context. The case-based MCQ Generator, trained through the use of GPT Builder, allows test developers to generate case-based MCQs easily. The benefits include enhanced efficiency in MCQ generation and the creation of contextually relevant questions surpassing standard ChatGPT capabilities. As Kiyak and Kononowicz demonstrate, researchers in language assessment can also develop test item generators with ChatGPT for various assessments in their field.

Shin and Lee (2023) assessed ChatGPT's potential in producing second language assessment materials comparable to those crafted by human experts. For human-made test materials, they used five reading passages and multiple-choice questions extracted from the English section of South Korea's College Scholastic Ability Test (CSAT), and AI-made test materials were generated using ChatGPT. For the study, a Likert-scale and open-ended survey was administered to 50 pre- and in-service teachers to measure their perceptions of the readings and testing elements. The findings showed that although the CSAT and ChatGPT-generated readings were perceived similarly in their natural flow and expressions, the CSAT readings were considered to have more appealing multiple-choice options and better quality in testing items. Through the study, the researchers suggest that ChatGPT has the potential to assist EFL teachers in generating reading passages and testing items, significantly reducing their workload. Additionally, they suggest that teachers should actively participate in revising the generated materials, considering the current limitations of ChatGPT.

Intrigued by the valuable aspects of ChatGPT, specifically about the test item generation, this study has been designed to investigate whether ChatGPT can function as a helpful tool for an instructor designing a TESOL achievement test in the field. This study examines whether an AI-assisted test form can parallel the human-made achievement test form.

Methodology

Participants and setting

The participants of this study were 43 students from a TESOL theory course titled *Materials and Methods in ELT.* The course, taught by the researcher, was offered in the fall semester of 2023 at a women's university in Seoul, South Korea. It aimed to provide third-year English-major students with knowledge and understanding of second-language teaching methods and materials. All 43 of the registered students participated in this project by taking the final test, which was a planned part of their educational curriculum. The students were informed about and consented to the use of the test data for research purposes. The test data were recorded in a way that ensured participants

could not be identified directly or through any related identifiers. Among the participants, thirty-seven participants were English majors, predominantly juniors, with a few seniors and sophomores. Additionally, six students were double majoring in English.

Alongside the 43 students, 3 female professors agreed to participate in analyzing the test content for parallel forms. All three raters have expertise in second language acquisition theories and English language teaching materials courses, with two holding master's degrees and one a doctoral in TESOL. Each individual possesses over 20 years of teaching and assessing experience at their respective educational institutions.

Instrument

Final test with 20 existing items and 20 new items

The final test for this study, administered in the fall semester of 2023, consisted of 40 items from two test forms. One form, containing 20 test items from the final test for the course administered in the fall semester of 2021, is dubbed Test A. The other form, comprising an additional 20 new test items from a parallel form created specifically for this study, is named Test B. Test A, the 2021 fall semester final test, included 10 true–false items and 10 multiple-choice items designed to assess students' knowledge in the course, covering contents related to teaching English listening, speaking, writing, and integrated skills. The 20 new test items for Test B were developed to reflect the test specifications of each item in Test A.

Item generation with the assistance of Chat-GPT

Item generation for Test B was conducted with the assistance of ChatGPT 3.5. In parallel and based on Test A, the 20 new items in Test B also consisted of 10 true–false and 10 multiple-choice items.

For true–false test items, AI-generated test items required almost no revision when prompted to create multiple true or false statements based on the sample statement derived from each existing test item. The test designer merely selected a statement from the various statements that ChatGPT provided. Consequently, designing 10 true–false items using ChatGPT was relatively simple, primarily involving the test designer choosing one option among several options generated by the AI. This efficiency can be attributed to the straightforward format of true–false questions and ChatGPT's capability to offer multiple options (see Fig. 1).

On the other hand, the multiple-choice test items required more complex procedures than the true–false test items due to the unsatisfactory quality of those generated by AI. Thus, each multiple-choice test item design necessitated additional steps involving further manipulation by the researcher.

Initially, distractors were created. ChatGPT was prompted to generate multiple statements based on the specific content covered and assessed in each test item from the 2021 fall semester final test. For example, to design a test item inquiring about "the characteristics of unskilled and skilled listeners," the AI was given the following prompt: Based on the following information: (Insert class contents about *unskilled* listeners), design 10 statements similar to the following statement: *"Unskilled listeners often ignore previous knowledge and context due to mishearing*." Subsequently, the AI provided 10 statements describing the characteristics of unskilled listeners suitable for use as distractors.

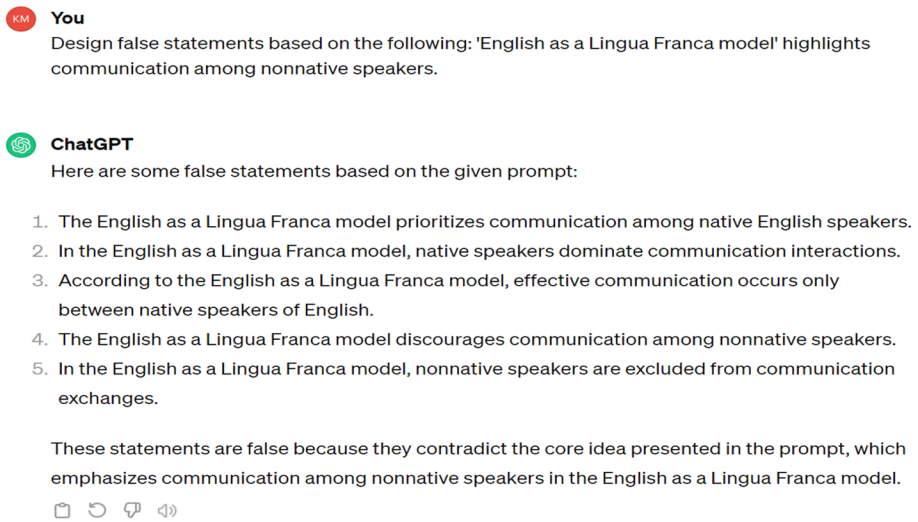

Fig. 1 Sample prompt and ChatGPT's true–false question output

Next, keys were generated. The AI tool was prompted to design multiple parallel keys referencing a key from the same multiple-choice item from the 2021 fall semester final test. For example, the following prompt was given: Based on the information provided (Insert class contents about *skilled* listeners), design multiple statements similar to the following: *"Skilled listeners seek contextual cues when there are problems in decoding."*

The researcher finally developed a multiple-choice test item by selecting the three best distractors from 10 statements about *unskilled* listeners and one key from a few statements describing *skilled* listeners. In this manner, each multiple-choice item was designed by reviewing several provided distractors and keys, making minimal changes if necessary.

Through this iterative process, 10 multiple-choice items were constructed through human-AI collaboration. While this method of designing multiple-choice test items was more time-consuming than creating true–false questions, it was perceived by the researcher/test designer to be less burdensome and laborious than previous test design procedures without using an AI tool.

Procedure

Raters' parallel form analyses of the test content

Before test administration, three independent TESOL professors reviewed 20 sets of paired questions, each comprising one item from the newly generated parallel form (Test B) and one from the existing test form (Test A). The three raters were asked to independently complete a questionnaire with 20 Likert-type items to investigate whether the two sets of 20 test items in each test form were designed to measure identical test specifications for specific TESOL contents. If each test item in Test A is judged to be parallel with its corresponding test item in Test B, the raters were instructed to choose the option 'strongly agree' and assign it a rating of 5, indicating the highest level of parallelism. Otherwise, they were to select the option 'strongly disagree' and assign it a rating of 1, denoting the lowest level of parallelism, on a Likert-type scale ranging from 1 to 5. They were

not informed in advance that one of the two tests was designed with the assistance of ChatGPT.

Test administration

All 40 items from Tests A (20 items from the 2021 fall semester final test) and B (20 new items generated with the assistance of ChatGPT) were entered into the university's Learning Management System (LMS) for computer-based test administration during class time. The 40 test items and test item distractors were scrambled in random order for each participant, and the LMS system did not allow the test takers to reenter the system once they finished and submitted their test.

The test was administered to the participants for an hour using the school's computers in a classroom large enough for about 80 students. Each of the 43 students was assigned a designated seat with a computer and keyboard, with the seat next to each left empty. The researcher served as a proctor, monitoring the process at the back during the test administration. Students were permitted to leave the room only upon completion of the test.

Data analysis

Data analysis was conducted using EXCEL and SPSS 29 with the R 4.3.2 extension. Before data collection, the internal consistency of Test A was examined using student test results from the 2021 fall semester final test. The calculated Cronbach's alpha value for the scale was 0.79, exceeding the minimum internal consistency coefficient threshold of 0.70 (Adadan & Savasci, 2011). Due to test security concerns, Test B could not undergo piloting, and therefore, its reliability was not assessed before administration.

To address the first research question, parallelism analyses were conducted on the content of the two tests. To investigate the second research question, which pertained to the parallelism of the two test forms, comparative analyses were performed on student test scores and item analyses. Initially, Fleiss' kappa was calculated to determine the level of agreement among the three raters' Likert-scale data, assessing the degree of parallelism for each set of 20 items in both test forms based on their content.

Subsequently, reliability and descriptive analyses were performed using student test results, and a bioequivalence test was utilized for the parallelism analysis of the two test forms to answer the second research question. Unlike the conventional null hypothesis, which aims to reject the null hypothesis indicating no difference between the two groups, this study employed a bioequivalence test to specifically reject the null hypothesis, suggesting that the mean of Test A is not equivalent to the mean of Test B:

$$H_0 : \mu_1 \neq \mu_2 \text{vs.} H_{1:} \mu_1 = \mu_2$$

To support H_1: $\mu_1 = \mu_2$, Schuirmann's (1987) two one-sided tests of equivalence of paired samples (TOST-P) were employed. In spite of the small sample size ($N = 43$) and the skewed distribution ($W = 0.94$, p value < 0.05), as the sample sizes of 30 are typically seen to be sufficient for the central limit theorem (CLT), TOST-P seemed to be an appropriate test for this study. Thus, to examine if the two test forms (Tests A and B) are parallel, TOST-P was calculated under the following hypotheses:

$$H_0 : \mu_D < -\Delta_{\mathrm{L}} \text{or } \mu_{\mathrm{D}} > \Delta_{\mathrm{U}} \text{vs.} H_{1:} - \Delta_{\mathrm{L}} < \mu_{\mathrm{D}} < \Delta_{\mathrm{U}}$$

For the upper and lower equivalence range, one (−1, 1) was used as simulated in Mara and Cribbie (2012). Furthermore, item analyses were conducted for each test form following established methodologies from the literature (Malau-Aduli et al., 2012; Precht et al., 2003).

Classical test theory (CTT) was employed to calculate item difficulty levels and item discrimination indices. Despite the limitations associated with CTT, such as difficulties in interpreting changes in scores over time and reliance on sample characteristics, using CTT in this study appears to be a reasonable choice. This decision is based on the administration of two achievement test forms to the same students at one time. Furthermore, given that classroom teachers can easily utilize standard statistical software for conducting analyses, CTT remains widely used in education and psychology due to its practicality (Ayanwale et al., 2022; De Champlain, 2010).

Results

Comparative analysis of tests A and B

To address the first research question, which investigates whether the contents of the 20 test items from each test form mirror each other, the three raters' assessment of the 20 sets of items demonstrated fair agreement with statistical significance ($p = 0.006, < 0.05$) (refer to Table 1). The computed Fleiss' kappa value was 0.35, with a 95% confidence interval (CI) ranging from 0.100 to 0.606. All three raters strongly agreed that most of the test items from the two forms (Tests A and B) were parallel, with mean scores of 4.85, 4.95, and 4.85 out of a maximum score of 5, respectively.

The answer to the second research question, investigating whether the two forms were parallel, was affirmative. Descriptive statistics for the two tests are presented in Table 2.

Table 1 Fleiss' Kappa analysis of three raters' responses

	Mean (SD)	Kappa	*z*	*p* value
Rater 1	4.85 (0.37)	0.35	2.74	0.006
Rater 2	4.95 (0.22)			
Rater 3	4.85 (0.37)			

Table 2 Descriptive statistics and reliability indices of tests A and B ($N = 43$)

	Test A (Human)	Test B (AI-Human)
# of items	20	20
Mean (SD)	14.19 (3.57)	14.53 (3.67)
Median	16.00	16.00
Maximum score	20	20
Minimum score	5	7
Skewness	−.78	−.50
Kurtosis	.16	−.76
Reliability (Cronbach's alpha)	0.73	0.78

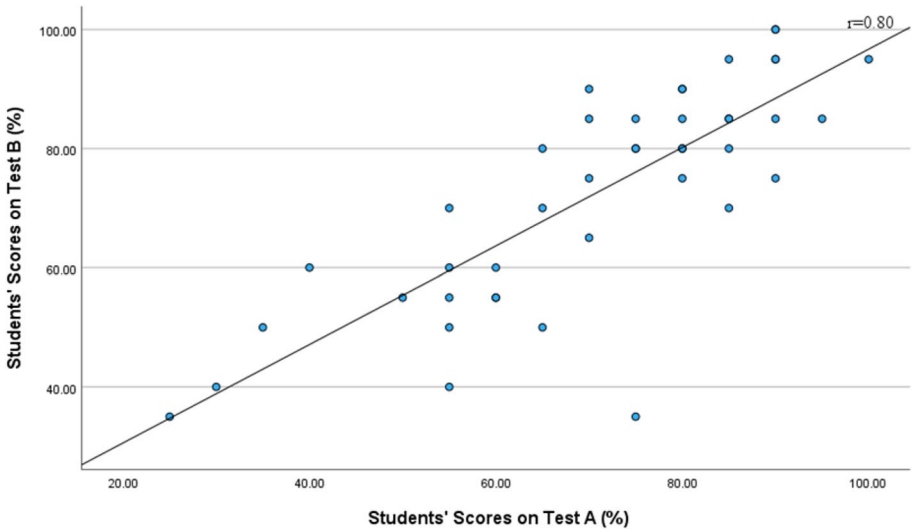

Fig. 2 Correlation between students' scores on tests A and B

As shown in Table 2, the mean scores were similar (14.19 for Test A and 14.53 for Test B), falling within the standard deviation ranges of 3.57 and 3.67, respectively. Both tests exhibited negatively skewed distributions, with values of −0.78 and −0.50 for Tests A and B, suggesting they were relatively easy for the students. Although Test B appeared more symmetrical than Test A, with a skewness value closer to 0 (−0.50), both tests can be considered close to symmetrical, as a skewness value between −1 and +1 is considered indicative of excellent data distribution (Hair et al., 2022). The kurtosis results for both tests were 0.16 and −0.76 for Test A and B, respectively, falling within the range of −1 to +1, indicating a near-normal distribution (Hair et al., 2022), albeit Test B displayed a flatter distribution compared to Test A. The calculated Cronbach's alpha values for each scale were 0.73 for Test A and 0.78 for Test B, satisfying the minimum internal consistency coefficient of 0.70 (Adadan & Savasci, 2011), with Test B displaying slightly higher reliability than Test A.

The paired samples *t*-test displayed that the mean difference between the two test forms, Test A ($n=43$, $M=14.19$, SD = 3.57) and Test B ($n=43$, $M=14.53$, SD = 3.67), is not significantly different with a p value of 0.319 ($p<0.05$). The two one-sided tests of equivalence of paired samples (TOST-P) examining the equivalence of the two test forms rejected both of the two null hypotheses, Δ_L, $t(42)=1.88$, $p=0.033$ and Δ_U, $t(42)=-3.90$, $p<0.001$ (<0.05). Thus, the data analyses indicated that the two tests were equivalent since the mean difference is not statistically different, and the 95% CI falls within the equivalence interval of ±1 with a weak effect size of $r=0.154$. Therefore, the data analyses indicated that the two tests were parallel.

As expected, the student scores on Tests A and B were closely related. As illustrated in the scatter plot graph in Fig. 2, a strong correlation was found between the student scores on Tests A and B, with $r(42)=0.80$, $p=0.001$ (<0.01).

Table 3 Item analysis for tests A and B ($N = 43$)

Test form	Test A (human)	Test B (AI-human)
Task type	*True–false test task*	
# of true–false items	10	10
Difficulty level (%)		
Easy (73–100)	3 (30)	5 (50)
Medium (28–72)	7 (70)	5 (50)
Difficult (0–27)	0	0
Mean difficulty % (SD)	67.91 (0.09)	76.74 (0.15)
Items with discrimination indices (%)		
Poor (< 0.19)	1 (10)	2 (20)
Fair (0.19–0.29)	0 (0)	2 (20)
Good (0.3–0.39)	3 (30)	2 (20)
Excellent (> 0.40)	6 (60)	4 (40)
Mean discrimination index (SD)	0.49 (0.17)	0.38 (0.24)
Task type	*Multiple-Choice Test Task*	
# of multiple-choice items	10	10
Difficulty level (%)		
Easy (73–100)	5 (50)	5 (50)
Medium (28–72)	5 (50)	5 (50)
Difficult (0–27)	0	0
Mean difficulty % (SD)	73.95 (0.15)	68.60 (0.18)
Items with discrimination indices (%)		
Poor (< 0.19)	4 (40)	1 (10)
Fair (0.19–0.29)	1 (10)	0 (0)
Good (0.3–0.39)	0 (0)	1 (10)
Excellent (> 0.40)	5 (50)	8 (80)
Mean discrimination index (SD)	0.37 (0.30)	0.53 (0.26)
Total # of items	*20*	*20*

Item analyses of tests A and B

For further analysis of each test item employed in this study, item analyses were conducted on 20 test items in each test form (see Table 3). Despite some differences, the item analyses of the two test forms displayed little deviation.

Regarding the 10 true–false test items in Test A, designed by the current researcher in 2021, and the parallel 10 in Test B, generated by Chat-GPT and selected by the researcher, the mean difficulty of the 10 true–false questions in each test form showed little deviation from the other (67.91% vs. 76.74%); however, that of Test B was closer to 85.00%, the ideal difficulty level for the true–false test task (Lord, 1952). Table 3 illustrates three true–false items were easy on Test A, whereas five easy items were observed on Test B. Seven on Test A and five on Test B were of medium difficulty levels for the remaining items on each test.

In terms of true–false item discrimination, on Test A, 60% (6 out of 10 total true–false items) had excellent discrimination indices with a mean discrimination index of 0.49 (> 0.40); on Test B, 40% (4 out of 10) had good discriminatory power with a mean discrimination index of 0.38 (0.3–0.39). One item (10%) on Test A had low discriminatory indices (< 0.19) and needed to be discarded, and two items (20%) on Test B needed to be deleted. Although items on Test A were slightly more effective than

Table 4 Distractor analysis for tests A and B ($N = 43$)

	Test A (human)	Test B (AI-human)
# of multiple-choice items	10	10
# of distractors assessed	40	40
Distractors with frequency = 0%	4 (10%)	3 (7.50%)
Distractors with frequency < 5%	9 (22.50%)	7 (17.50%)
# of functioning distractors (%)	27 (67.50%)	30 (75.00%)
Functioning distractors per item mean (SD)	2.70 (0.78)	3 (0.77)
Functioning distractors per item *n* (%)		
None	0 (0)	0 (0)
One	1 (10)	0 (0)
Two	2 (20)	3 (30)
Three	6 (60)	4 (40)
Four	1 (10)	3 (30)

those on Test B, considering the small number of the assessed items, the true–false questions on both tests were roughly practical.

Based on the item analysis performed on the multiple-choice test items, which comprised 10 researcher-designed items in Test A and 10 in Test B, where the distractors and keys were generated by Chat-GPT and selected and arranged by the current researcher, the mean difficulty of the test items for each test was similar, with 73.95% for Test A and 68.60% for Test B. Although the mean difficulty levels of both tests were close to 74%, the ideal difficulty level for the four-response multiple-choice test (Lord, 1952), Test A's difficulty level was marginally closer to Lord's suggested level.

The results of the multiple-choice item analyses demonstrate that the test items on both test forms yielded similar outcomes, with Test B displaying better discriminating indices than Test A. On Test B, 80% (8 out of 10) had excellent discriminatory power with a mean discrimination index of 0.53 (> 0.40). In contrast, on Test A, only 50% (5 out of 10) displayed such excellent discrimination indices with a mean discrimination index of 0.37 (< 0.40). This result may be attributed to the fact that the items on Test A served as references for creating the items on Test B. Constructing test items based on an existing sample can often be easier than creating them from scratch. Additionally, this outcome suggests that technology may potentially be more efficient when collaborating with human intelligence and skills.

Distractor analyses of the 10 multiple-choice items for each test form were performed to examine how well each option functions for the quality of each item. Forty options, including 30 distractors and 10 keys in 10 test items on Test A, were assessed, and the corresponding 40 options on Test B were examined (see Table 4).

As illustrated in Table 4, similar patterns were observed in both test forms, with Test B having slightly more functioning distractors. In Test A, 27 distractors (67.50%) were functional, compared to 30 (75%) in Test B. Four distractors (10%) were not chosen by any examinee in Test A, compared to three (7.5%) in Test B. The mean number of functioning distractors per item was close, with 2.70 in Test A and 3 in Test B. Most items in both tests had more than two functioning distractors, with the largest number having three functional distractors (60% for Test A and 40% for Test B). However,

Test A had one item with only one functioning distractor, whereas no such item was observed in Test B. Moreover, Test B had three items (30%) with four fully functioning distractors, while only one item (10%) in Test A was found with four functioning distractors. Overall, the distractor analyses of the two tests illustrate a similar pattern to some extent; however, Test B had more items with well-functioning distractors.

Discussion

This study aimed to examine the efficiency of ChatGPT in designing parallel test items alongside those created by humans. The study employed item content analysis and classical test theory (CTT) to examine the parallelism of the two test forms: an AI-human-made test form (Test B) and a human-made test form (Test A).

In the initial inquiry, examining the content parallelism of the two tests, the analysis conducted by three experts indicated affirmative findings. All three raters strongly agreed on the parallelism of most test items across forms, reflected in their high mean ratings. Moreover, the examination revealed fair agreement among the raters, suggesting a statistically significant level of agreement.

The finding is in the same vein as Shin and Lee (2023), who presented the blind test results of pre-and in-service English teachers in South Korea responding to both the ChatGPT-generated reading passage and its counterpart on the College Scholastic Ability Test. In Shin and Lee's study, both pre-and in-service teachers displayed no difference between the two types of passages regarding naturalness in flow and expressions, as they strongly agreed with the Likert-scale items.

Based on the findings of both studies, it appears that providing a sample, as in the current research and Shin and Lee's study, proves to be very helpful in generating quality outcomes from ChatGPT. In that sense, generating parallel test items for large test item banks appears to be a solid idea for employing the AI tool.

The study also investigated whether the two test forms were parallel in assessing students' performances. The statistical analysis indicated the AI-human-made and human-made test forms were equivalent within the equivalence interval. Moreover, despite some observed variations, the item and distractor analyses suggested that the AI-human test items were of sufficient quality to be used as an alternative to human-made achievement test items. These findings are consistent with those of Kiyak, Coşkun, Budakoğlu, and Uluoğlu (2024) and Shin and Lee (2023), demonstrating ChatGPT's capacity to generate test items.

ChatGPT 3.5 was particularly useful for designing true–false test items. When a sample true statement was entered requesting multiple true statements, ChatGPT successfully generated true statements similar to the sample. When false statements were needed, the same conditions were provided for the true statements, but multiple false statements were requested, and ChatGPT provided usable false statements.

However, the ChatGPT could not generate well-functioning multiple-choice test items and thus needed multiple steps involving human intelligence, as detailed in the methodology section. Hence, for designing the final test, human intelligence was involved in selecting statements for the true–false test items and selecting and organizing options for multiple-choice test items. Therefore, considering the current capacity of

ChatGPT 3.5 to design quality test items, human intelligence seems to still be required, and this finding aligns with those from previous studies emphasizing that AI tools are not infallible and require human intelligence (Aryadoust et al., 2024; Shin & Lee, 2023).

Nonetheless, despite necessitating several steps involving human intelligence, the AI-assisted test construction was less demanding for the researcher/test developer than the process without the AI because it provided multiple options instantly, resulting in significantly less time spent making multiple-choice items than designing them from scratch. In understanding the educational practices where instructors sometimes use the exact test items on a test that were utilized before in spite of the negative consequences (Wellberg, 2023), this study seems to be a case in point illustrating ChatGPT can be helpful in the creation of a parallel test form for an achievement test. With specific test specifications or sample test items provided, ChatGPT is expected to work as a capable assistant for instructors. Although there are some concerns regarding the potential disruptions of traditional assessment practices due to AI-driven tools (e.g., Ibrahim, 2023), the current research and data analyses suggest that not utilizing ChatGPT for designing a parallel test form can be seen as inefficient, akin to taking the long way round instead of the quick and efficient way. Consequently, this study reaffirms previous findings that AI applications such as ChatGPT can alleviate teachers' workloads (Baker et al., 2019) and save time (Koltovskaia, 2020).

Conclusion

This study examined the effectiveness of ChatGPT in designing parallel test items compared to those created by humans. The findings from the first inquiry, which investigated whether the two tests covered parallel content areas, yielded positive results. In the second inquiry, which examined the comparability of the two tests in assessing student performance, the statistical analysis indicated the AI-human-made and human-made test forms were equivalent within the assigned equivalence interval.

The study's findings illustrate the efficacy of employing ChatGPT to create final test items for a university TESOL theory class called *Materials and Methods in ELT*. The AI-assisted test items seem to be a practical solution when instructors in the field are often challenged with repeatedly creating new test items every semester or year. The current capacity of ChatGPT 3.5 required human involvement to some extent since the researcher was involved in selecting or both selecting and organizing options for the true–false and multiple-choice test items, respectively. However, creating new test items for achievement tests with AI assistance was, at least for the researcher, less challenging, time- and effort-consuming than without one.

Despite the current design's positive outcomes and methodological triangulation, the study has several vulnerabilities. First, the study is limited in selecting the participants and the sample size. Since the study employed a convenient sampling method of using one of the researcher's classes, which involved only 43 students, the results of the current research should not be generalized to other populations. Thus, a more systematic research design involving a larger sample size and random sampling is expected to be followed.

Moreover, since the class was taught strictly according to a fixed syllabus covering TESOL theory, this study was limited to item content analysis. It focused on

investigating the contents according to test specifications without further validity analysis of test constructs. A replicate study using other skills tests with specific constructs such as English reading comprehension tests, French-speaking skills tests, or English grammar tests would render more systematic research designs.

In addition, as a small-scale analysis employing a small number of participants taking the two test forms simultaneously, this study solely relied on classical test theory (CTT) for item analysis. The item difficulty in the classical model can be unstable depending on the sample and the test form, and more importantly, the classical model can produce more measurement errors compared to item response theory (IRT). Since the current study is limited in scope due to the small sample size and the small-scale analysis of the final test result, conducting more rigorous analyses on a larger population of over 100 individuals with a medium-stakes examination using the item response theory (IRT) model would yield more insightful educational implications with more substantial datasets. In addition, considering that the equivalence test result can vary based on the assigned range of equivalence, more studies with diverse equivalence ranges would yield more powerful analyses. With AI technology rapidly advancing, more systematic research analyses on AI-generated test items are expected to emerge soon.

Given the recent emergence of ChatGPT, there remains a dearth of research studies exploring its application in the field of language learning and assessment. Through the comparative analysis of the two test forms, this study demonstrates AI's potential as a convenient tool for assisting teachers in designing test items, particularly in creating parallel items to existing ones.

For future studies, research could explore additional factors influencing the efficacy of ChatGPT in test item creation and its integration into broader educational practices. Despite limitations in sample size, participant selection, and reliance on CTT, this study's findings offer implications for both theory and practice, highlighting the potential of ChatGPT to facilitate test item creation processes, reduce teacher workload, and enhance efficiency in educational assessment.

Abbreviations

AI	Artificial Intelligence
ChatGPT	Chat Generative Pre-Trained Transformer
CSAT	College Scholastic Ability Test
CTT	Classical Test Theory
IRT	Item Response Theory
LMS	Learning Management System
MCQ	Multiple-Choice Question
NLP	Natural Language Processing
SPSS	Statistical Package for the Social Sciences
TESOL	Teaching English to Speakers of Other Languages
TOST-P	Two One-Sided Tests of Equivalence of Paired Samples

Acknowledgements
The author would like to thank all the students and professors who kindly participated in this study. Furthermore, the author expresses gratitude to the journal editor and reviewers for their feedback and support.

Author's contributions
KO initiated and conducted the research autonomously, collecting data, performing statistical analyses, drafting the manuscript, and approving the final version.

Funding
The research benefited from the support provided by the research fund of Dongduk Women's University.

Availability of data and materials
The data used during the current study are available from the author upon reasonable request.

Declarations

Competing interests
The author affirms the absence of any competing interests.

Received: 13 March 2024 Accepted: 24 May 2024
Published online: 07 June 2024

References

Abida, F. I. N., Kuswardani, R., Purwati, O., Rosyid, A., & Minarti, E. (2023). Assessing language proficiency through AI chatbot-based evaluation. In *Proceedings of the International Conference on Islamic Civilization and Humanities* (Vol. 1, pp. 138–145). Retrieved March 9, 2024 from https://proceedings.uinsby.ac.id/index.php/iconfahum/article/view/1230.

Adadan, E., & Savasci, F. (2011). An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education, 34*(4), 513–544. https://doi.org/10.1080/09500693.2011.636084.

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*(2024), 100204. https://doi.org/10.1016/j.caeai.2024.100204.

Ayanwale, M., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: the status of the framework at the examination council of Lesotho. *International Journal of Learning, Teaching and Educational Research,21*(8), 384–406. https://www.ijlter.org/index.php/ijlter/article/view/5676.

Baker, T., Smith, L., & Anissa, N. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges*. Nesta Foundation. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf.

Benali, A. (2021). The impact of using automated writing feedback in ESL/EFL classroom contexts. *English Language Teaching, 14*(12), 189–195. https://doi.org/10.5539/elt.v14n12p189.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020, December 6–12). *Language models are few-shot learners. Paper presented at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (pp. 1877–1901). Vancouver, Canada. Retrieved March 9, 2024 from https://doi.org/10.5555/3495724.3495883.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessment in medical education. *Medical Education,44*(1), 109–117. https://pubmed.ncbi.nlm.nih.gov/20078762/.

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

Gehringer, E. (2004). Reuse of homework and test questions: When, why, and how to maintain security? In *Proceedings of the 34th Annual Frontiers in Education Conference* (pp. S1F/24-S1F/29). Retrieved March 9, 2024 from https://doi.org/10.1109/fie.2004.1408702.

Gierl, M. J., Lai, H., & Turner,. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Thousand Oaks: Sage. https://www.pls-sem.net/pls-sem-books/a-primer-on-pls-sem-3rd-ed/.

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation, 5*(1), 37–45. https://jeti.thewsu.org/index.php/cieti/article/view/103.

Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "the terminator versus the machines." *Language Testing in Asia, 13*(1), 46. https://doi.org/10.1186/s40468-023-00260-2.

Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education, 206*, 104898. https://doi.org/10.1016/j.compedu.2023.104898.

Kiyak, 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, 2024*. https://doi.org/10.1007/s00228-024-03649-x.

Kiyak, Y. S., & Kononowicz, A. A. (2024). Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. *Medical Teacher*. https://doi.org/10.1080/0142159X.2024.2314723.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450. https://www.sciencedirect.com/science/article/abs/pii/S1075293520300118.

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*(7), 5614. https://doi.org/10.3390/su15075614.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2019). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y.

Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika, 18*, 181–194. https://doi.org/10.1007/BF02288781.
Malau-Aduli, B. S., Walls, J., & Zimitat, C. (2012). Validity, reliability and equivalence of parallel examinations in a university setting. *Creative Education, 3*, 923–930. https://www.scirp.org/journal/paperinformation?paperid=23559.
Mara, C. A., & Cribbie, R. A. (2012) Paired-samples tests of equivalence. *Communications in Statistics - Simulation and Computation, 41*(10), 1928–1943. https://www.tandfonline.com/doi/abs/10.1080/03610918.2011.626545.
Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P. C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challnges. *BioData Mining,16*, 20. https://doi.org/10.1186/s13040-023-00339-9.
Nagarhalli, T. P., Vaze, V., & Rana, N. K. (2020). In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). A review of current trends in the development of chatbot systems* (pp. 706–710). Coimbatore https://www.researchgate.net/publication/340893457_A_Review_of_Current_Trends_in_the_Development_of_Chatbot_Systems.
Pranav, D. S., Mutreja, M., Punj, D., & Chawla, P. (2022). Natural language processing in chatbots. In *Emerging Technologies in Data Mining and Information Security* (pp. 87–98). Proceedings of IEMIS 2022, 3. Institute of Engineering & Management, Kolkata, India. Retrieved March 9, 2024 from https://doi.org/10.1007/978-981-19-4193-1_9.
Precht, D., Hazlett, C., Yip, S., & Nicholls, J. (2003). *Item analysis user's guide*. International Database for Enhanced Assessments and Learning
Pugh, D., Champlain, A. D., Gierl, M., Lai, H., & Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical Teacher, 38*(8), 838–843. https://doi.org/10.3109/0142159X.2016.1150989.
Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of the traditional assessment in higher education. *Journal of Applied Learning & Teaching, 6*(1), 242–263. https://doi.org/10.37074/jalt.2023.6.1.9.
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://pubmed.ncbi.nlm.nih.gov/3450848/.
Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics, 8*, 247–263. https://doi.org/10.1162/tacl_a_00310/96485/Machine-Learning-Driven-Language-Assessment.
Shewale, R. (2024, January 12). *ChatGPT Statistics—User Demographics*. DemandSage. Retrieved March 9, 2024 from https://www.demandsage.com/chatgpt-statistics/.
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://hdl.handle.net/10125/73530.
Swiecki, Z., Khosravi, H., Chen, G. L., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Ga˘sevi´c, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075.
Taecharungroj, V. (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on twitter. *Big Data Cognitive Computing, 7*(1), 35. https://doi.org/10.3390/bdcc7010035.
Thao, N. T. P. (2023). The application of ChatGPT in language test design: The what and how. In *Proceedings of the AsiaCALL International Conference (AsiaCALL2023)* (pp. 104–115). Retrieved March 9, 2024 from https://asiacall.info/proceedings/index.php/articles/article/view/84.
Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 163–173). Association for Computational Linguistics. Retrieved March 9, 2024 from https://doi.org/10.5555/2390384.2390404.
Wellberg, S. (2023). Teacher-made tests: Why they matter and a framework for analysing mathematics exams. *Assessment in Education: Principles, Policy & Practice, 30*(1), 53–75. https://www.tandfonline.com/doi/full/10.1080/0969594X.2023.2189565.

Publisher's Note