Research Article

The use of ChatGPT in assessment

Mehmet Kanık [1*]

[1]Final International University, Faculty of Educational Sciences, English Language Teaching Program, Girne, North Cyprus

Abstract: ChatGPT has surged interest to cause people to look for its use in different tasks. However, before allowing it to replace humans, its capabilities should be investigated. As ChatGPT has potential for use in testing and assessment, this study aims to investigate the questions generated by ChatGPT by comparing them to those written by a course instructor. To investigate this issue, this study involved 36 junior students who took a practice test including 20 multiple-choice items generated by ChatGPT and 20 others by the course instructor, resulting in a 40-item test. Results indicate that there was an acceptable degree of consistency between the ChatGPT and the course instructor. Post-hoc analyses point to consistency between the instructor and the chatbot in item difficulty, yet the chatbot's results were weaker in item discrimination power and distractor analysis. This indicates that ChatGPT can potentially generate multiple-choice exams similar to those of the course instructor.

1. INTRODUCTION

Swiecki et al. (2022) criticize standard assessment paradigms for being onerous, discrete, uniform, antiquated, and lacking authenticity. They propose that artificial intelligence (AI) can offer solutions to these challenges. In a review article on the use of AI in student assessment, González-Calatayud et al. (2021) argue that AI technologies remain underutilized in education due to users' lack of knowledge. However, within the past few years, there have been discussions on the impact of AI language models with the emergence of ChatGPT, a chatbot released by a company named OpenAI (chat.openai.com). This interest has also led to a surge in research studies in education, primarily focusing on language learning (Crompton & Burke, 2023).

Nevertheless, ChatGPT came with concerns and controversies, especially within the field of education. One of the initial reactions was of the negative kind as reports revealed that students had ChatGPT or other AI models to write projects and homework assignments for them. However, these language models may also offer some potential benefits and uses. For instance, Okonkwo and Ade-Ibijola (2021) identified several possible uses of chatbots in education including teaching, learning, and assessment. In Crompton and Burke's recent review (2023), themes such as assessment/evaluation, prediction, AI assistance, intelligent tutoring systems,

*CONTACT: Mehmet KANIK ✉ mehmetkanik@gmail.com ⊡ Final International University, Faculty of Educational Sciences, English Language Teaching Program, Girne, North Cyprus

and student learning management emerged as common applications of AI in education. This underscores the potential of AI language models like ChatGPT in education. Yet, before making use of such technologies, it is crucial to scrutinize their use, supported by evidence, as they may not always produce satisfactory or accurate content (van Dis et al., 2023). Therefore, this study attempts to investigate the use of ChatGPT in test preparation and assessment.

1.1. Literature Review

Gardner et al. (2021) state that Page (1966) "foresaw a time in the future when natural language processing (NLP) would achieve the technical maturity to enable machines to learn and understand how to assess the existence of the many complex trins in human writing" (p. 1208). Gardner et al. (2021) elaborate on this idea, asserting that machines can assess students on their knowledge of the content if the machine is trained on that content and trained to ask questions. To some extent, Page's (1966) prediction has become a reality as AI technologies now possess such capabilities. They can even do more. ChatGPT, for instance, has great capabilities that can contribute to teaching and assessment in a variety of ways. In an article, for instance, Lo (2023) reviewed studies on ChatGPT and identified five key uses of it in teaching and assessment, ranging from generating course materials to performing language translation. Lo (2023) also suggests that students can use it in preparing writing assignments for assessment. They can draft papers and have ChatGPT evaluate them for errors, and then the students can finalize their papers. As such, the chatbot can act as a useful scaffolding tool. According to this review, instructors can have it generate assessment tasks and evaluate student performance. Assessment and evaluation emerged as the most common use of AI technologies in higher education, as revealed by Crompton and Burke's (2023) review, encompassing automatic assessment, test generation, feedback, online activity review, and the evaluation of educational resources. Formative assessment, automated scoring, and comparisons between AI and non-AI assessment methods are also central to the research on assessment (González-Calatayud et al., 2021).

In a more detailed look at the contributions AI can make to overcome the problems in the standard assessment paradigm, Swiecki et al. (2022) suggest such uses as automated assessment construction, AI-assisted peer assessment, writing analytics, electronic assessment platforms, stealth assessment, latent knowledge estimation, learning processes, computerized adaptive testing, virtual simulations to add authenticity and modernized digital assessment by incorporating computational media such as AI-supported word processing. As an AI tool, Halaweh (2023) highlights the time and effort ChatGPT helps save and compares it to other tools like search engines and spreadsheets that are used to help with searching for information, calculations, and organizing data without concern, which were tasks that people had to do without the assistance of technology. The researcher suggests that as there are no concerns with using these tools so should there be no concern with using ChatGPT's abilities to produce and edit texts by considering it as a tool to save time and effort.

Yet, there are obvious concerns about the ethicality of using ChatGPT as it is capable of producing texts quickly and can cause ethical issues when used to replace one's role as the writer of a text. Dowling and Lucey (2023) found, for example, that ChatGPT can produce articles that can go through a peer-review process as the three articles produced by ChatGPT got high ratings from the reviewers. If the authorship is falsely claimed, they suggest, then ethical issues ensue. ChatGPT poses some issues for the users as well. For example, it can rely on biased data, not having up-to-date information, and generate incorrect or fake information. It can also present issues to educators related to ethical concerns. It can lead students to be involved in plagiarism and have them bypass plagiarism detectors (Lo, 2023, p. 8). Mhlanga (2023), thus, suggests responsible and ethical uses of ChatGPT in education by highlighting factors including responsible AI use and educating students about it and its limitations, transparency in the use of ChatGPT, respect for privacy, accuracy of information, and the like.

Lo (2023) suggests that instructors can benefit from using ChatGPT as a valuable resource, as it helps in crafting course syllabi, teaching materials, and assessment tasks as long as issues related to the accuracy of the generated content are addressed. Al-Worafi et al. (2023) tried the feasibility of using ChatGPT for designing curriculum and syllabus, course content preparation, and writing exams. The chatbot got expert ratings from 50% to 92%. Overall, it could be suggested that ChatGPT can be a useful tool. One aspect that the researchers looked at was exam preparation and found that ChatGPT can be used for that purpose. The expert rating of appropriateness and accuracy of what ChatGPT produced was 70%. They caution, however, that the exams did not include all the learning outcomes. Other AI tools were used in studies to generate cloze tests and found that AI tools can enhance learning (Olney et al., 2017; Yang et al., 2021).

Regarding exam generation, Chen et al. (2018) mention two methods, rule-based and data-driven, used in automatic question generation, creating strong potential for AI use in education. They suggest that the rule-based method is prone to be influenced by the quality and quantity of rules developed by humans, which will be dependent on their knowledge, experience, and effort. They suggest, as an alternative, the use of data-driven methods which will not be dependent on human-generated rules. Their research with a data-driven method indicates the data set can affect the extent to which automatic question-generation methods can write quality items as their research shows that automatic question-generation methods did not perform well in a comprehensive data set.

Another aspect AI language models were used for was the a priori evaluation of the quality of the exams generated by humans. For example, Moore et al. (2022) utilized GPT-3 to evaluate the quality of the student-generated short-answer questions. Although their focus was on the extent to which students are able to generate quality test items, the results also indicated the use of GPT-3 in evaluating and assessing the content of students' work. They found, however, that GPT-3 matched human evaluation only for 40% of the questions. For the AI model, most of the questions were high quality as opposed to human experts who classified 68% of the questions as low quality. For GPT-3 this figure was only 9%. The researchers conclude that GPT-3 overestimated the quality of the questions. In assigning the items to the levels of Bloom's taxonomy, there was a disagreement between GPT-3 and human experts in 68% of the questions.

In another study, however, Moore et al. (2023) utilized GPT-4 along with human and automated rule-based methods in evaluating the quality of items by identifying item-writing flaws in multiple-choice items. They found that GPT-4 was able to identify 79% of the flaws identified by human annotators and matched 62% of the human quality evaluations. This may indicate that the more advanced language models become, the better they can perform pedagogical tasks, approximating the performance of experts. AI technologies have also been used in automated essay scoring and have been utilized commercially and in computerized adaptive testing both used commercially by testing companies like Pearson or ETS (Gardner et al., 2021). Thus, AI-based tools can automate traditional assessment by creating tests and automatically scoring them, eliminating some of the burden (Swiecki et al., 2022). Swiecki et al. (2022) list some challenges of AI-based assessment tools. They caution against directly accepting machine decisions and giving the responsibility to engineers with no contact with the students also causing a removal of accountability. They are also skeptical about eliminating the pedagogical role of assessment teachers may use to affect the teaching-learning process, limiting the process-based performance assessment. Another issue they raise is the data collection by AI technologies. These are quite valid concerns about AI-based technologies.

If AI tools like ChatGPT are used for creating exams or writing assessment tasks, exam validity and reliability become another concern because they are required qualities of any test (Thorndike & Thorndike-Christ, 2014). "Validity has to do with the degree to which test scores

provide information that is relevant to the inferences that are to be made from them" (Thorndike & Thorndike-Christ, 2014, p. 76) or put simply, measuring what we want to measure with it and usually focuses on content-, criterion-, and construct-related validity. Content validity is usually achieved by having an exam blueprint, or a table of specifications, which shows the content areas and cognitive processes involved and their respective weight in the test. Criterion and construct validation techniques may need correlation with other tests (Miller et al., 2013; Reynolds et al., 2009; Thorndike & Thorndike-Christ, 2014).

Reliability is defined as "the accuracy or precision of a measurement procedure" (Thorndike & Thorndike-Christ, 2014, p. 75) or "consistency or stability of assessment results" (Reynolds et al., 2009, p. 91) and it is essential for testing because the purpose of assessment is to make educational decisions and if the information to base the decisions on is not reliable, then the decisions are unlikely to be valid decisions (Reynolds et al., 2009; Thorndike & Thorndike-Christ, 2014) and essentially the test tests "nothing" (Thompson & Vacha-Haase, 2018, p. 231). The reliability of exams is usually measured by calculating a reliability coefficient by correlating the results of the same tests administered at different times, parallel forms of a test, two halves of a test, and scores awarded by different examiners (Reynolds et al., 2009). In all these approaches, the consistency between two sets of scores is at the focal point of measurement.

González-Calatayud et al. (2021) highlight that AI is mostly used for formative assessment. There do not seem to be studies focusing on its use in summative assessment by testing the applicability of tests generated by AI language models. To approach the issue more systematically, this study aims to analyze the results of an exam prepared by ChatGPT in tandem with the course instructor to better answer the question of whether ChatGPT can be used in test preparation by course instructors by running comparative post-hoc evaluations like reliability, item difficulty and discriminating power.

2. METHOD

This study employs a case study approach, incorporating both quantitative and qualitative research methods to provide a more in-depth analysis of the subject. To explore the quantitative aspect, correlation, paired-samples t-test and post-hoc analysis were utilized to examine the reliability between two tests and to examine various aspects of the test results. Qualitative data were also gathered and analyzed using content analysis. By combining quantitative correlational analysis with qualitative content analysis, this study aims to offer a rich, nuanced understanding of the case.

2.1. Context

The study was conducted at a private university in North Cyprus, which is an international university with a majority of international student population. The university has a faculty of educational sciences with both Turkish-medium and English-medium programs. English Language Teaching program, as well as all the other programs of the faculty, has a course on measurement and evaluation in education aiming to train student teachers on assessment and testing practices. The study is conducted within this class.

2.2. Participants

The participants were students enrolled in the said measurement and evaluation class offered as part of an undergraduate program in English Language Teaching. The class is a mandatory faculty class that all registered students should take. There were 44 students enrolled in the class. Thirty-six of them participated in the study by taking the review exam. One paper was eliminated for being incomplete as the student answered questions in one part of the exam which was mainly the instructor's questions and did not complete most of the questions written by ChatGPT. The participant profile is outlined in Table 1 below. The students come from Ivory Coast, Libya, North Cyprus, Russia, Türkiye, Turkmenistan, and Uzbekistan. Eighteen of the

participants were female while 17 were male. The mean age was 22.94, ranging from 21 and 28.

Table 1. *Participants.*

Nationality	N (35)	Age		Gender	
Türkiye	14	Mean	22.94	Female	18
Ivory Coast	7	Range	21-28	Male	17
Uzbekistan	6				
Libya	2				
North Cyprus	2				
Russia	2				
Turkmenistan	2				

2.3. Procedures

For the purpose of the study, a table of specifications including the content and learning outcomes of the said class was prepared. The table of specifications included 20 items distributed over the content of the class covered between the midterm exam and the final exam of the class. The same table of specifications was used also for the final exam of the class. The instructor of the class wrote 20 questions matching this table of specifications to ensure content validity. Then, the instructor pasted the content of the class lecture presentations into ChatGPT and asked the chatbot to write questions. A sample entry used, for example, reads "Using the following information, prepare a multiple-choice item on item analysis at Bloom's knowledge level". After ChatGPT produced 20 questions matching the same specifications, two sets were put together resulting in a 40-question multiple-choice test. Half of the students began with the instructor's questions, while the other half started answering the questions written by ChatGPT. Students were also asked to write their comments on the questions for their perception of the test and the questions.

After the administration of the test, each exam paper was given several scores: One total score, one score for the questions by the instructor, one score for the questions written by ChatGPT, two scores each for the odd and even-numbered questions written by the instructor, ChatGPT and combined total resulting in nine different scores. These scores were put into statistical software for analysis. The main methods of statistical analyses were correlation and reliability analysis. Item analysis procedures were also conducted for item difficulty, item discrimination power, and distractor effectiveness. Students' comments were analyzed qualitatively.

3. FINDINGS

The first analysis was calculating the internal reliability of the exam as well as the sections written by the instructor and ChatGPT. To calculate the internal consistency, the odd-numbered questions and the even-numbered questions were scored separately. This was done for the instructors' and ChatGPT's questions as well. The results of the analysis for the whole test yielded a score of .743, which is an acceptable value (Reynolds et al., 2009) as indicated in Table 2.

Table 2. *Split-half reliability analysis results.*

Test	N	Odd M (SD)	Even M (SD)	Spearman-Brown coefficient
Instructor's test	35	14.36 (4.59)	13.57 (5.60)	.636
ChatGPT's test	35	15.64 (3.94)	16.07 (4.21)	.636
Combined	35	30.00 (7.52)	29.57 (8.47)	.743

Split-half reliability analysis was also calculated for the instructor's test and ChatGPT's tests using the Spearman-Brown formula. The obtained coefficient for both the instructor's and ChatGPT's questions was .636, indicating a moderate internal consistency, understandably a bit lower than the combined test since the sample size goes down in split-half analysis and lower reliability coefficients can be acceptable for short tests (McCowan & McCowan, 1999). Table 2 shows these results.

After establishing an acceptable degree of internal consistency, parallel forms reliability was calculated between the instructor's test and ChatGPT's test. The coefficient calculated for the reliability between these two forms was .80, which points to a good degree of consistency. This finding is important as it indicates that ChatGPT can prepare tests that function parallel to a course instructor's test. The results are depicted in Table 3.

Table 3. *Consistency between the instructor's test and that of ChatGPT.*

Test	N	M (SD)	Spearman-Brown coefficient
Instructor's half	35	28.00 (8.71)	.80
ChatGPT's half	35	31.57 (6.91)	

The question of whether the instructor in question writes consistent exams is a question in point here. To establish that, the reliability between the instructor's two tests given at two different times of the semester was calculated. A reliability coefficient of .92 was achieved, indicating a good degree of reliability, as shown in Table 4

Table 4. *Reliability between two tests written by the course instructor.*

Test	N	M (SD)	Cronbach's Alpha
Test 1	35	68.17 (16.36)	.92
Test 2	35	55.71 (18.98)	

ChatGPT's ability to write tests consistent with the course instructor's tests indicates its utility in helping with testing and assessment. The mean scores of the tests indicate, however, that the instructor's version may have been more challenging. The paired sample t-test was run, and the results showed that the students achieved higher scores in ChatGPT's set (M=31.57, SD= 6.91) than in the instructor's set (M=28, SD=8.71), resulting in a significant difference as can be seen Table 5.

Table 5. *Paired samples t-test statistics.*

Test	N	M (SD)	t	df	p
Instructor's set	35	28.00 (8.71)	-3.204	34	.003
ChatGPT's set	35	31.57 (6.91)			

After the reliability analyses, item analysis procedures were followed to see if ChatGPT writes items with a good level of difficulty and discrimination power.

3.1. The Results of Item Analysis

The difficulty index of the items demonstrates similar results from the instructor's and ChatGPT's sets. As Table 6 depicts, 70% of the instructor's test items proved to have moderate levels of difficulty while 65% of ChatGPT's test items fell into the moderate difficulty range. Both sets of test items had two that were identified as difficult. In terms of the easy items, 20% of the instructor's and 25% of ChatGPT's items were in the easy range. These results indicate

that both the course instructor and ChatGPT write questions at a comparable degree of difficulty.

Table 6. *Difficulty index.*

	Instructor	ChatGPT
Easy	4 (20%)	5 (25%)
Moderate	14 (70%)	13 (65%)
Difficult	2 (10%)	2 (10%)

Another relevant analysis is the discrimination power of the items (Ebel & Frisbie, 1986). In this analysis, the ratio of the correct answers by lower achieving to those of the higher achieving students is calculated. The results indicate that 75% of the instructor's items are very good or reasonably good while only 50% of the items written by ChatGPT were good in terms of discrimination power. This indicates that ChatGPT fails to write items that can distinguish between the higher and the lower-achieving students. Table 7 outlines these results.

Table 7. *Discrimination index.*

	Instructor	ChatGPT
Very good	11 (55%)	8 (40%)
Reasonably good	4 (20%)	2 (10%)
Marginal item	3 (15%)	6 (30%)
Poor item	2 (10%)	4 (20%)

3.2. Distractor Analysis

For the 20 four-option test items, both the instructor and ChatGPT wrote 60 distractors in total. The expectation for the distractors is that they are to be selected by some students and selected by the low-achieving students more than the high-achieving students (Miller et al., 2013). According to the results of the analysis, 90% of the distractors written by the instructor were selected by some students, while only 80% of those written by ChatGPT were selected by some students. The number of the instructor's distractors that were selected by the lower group of students is 41, accounting for 71.6% of the total distractors. While this value is 34 for ChatGPT accounting for 56.6% of the distractors it wrote. In other words, 43.4% of the distractors written by ChatGPT were poor distractors as opposed to 28.4% of the instructor as shown in Table 8. This can indicate that ChatGPT may not be apt to write plausible distractors.

On the other hand, in this specific case, the instructor may sometimes be writing distractors that may be confusing, as 10% of the distractors were selected more by the upper group, which indicates an issue. On the other hand, only one distractor written by ChatGPT was selected by the upper group more than the lower group. Thus, ChatGPT may be clearer in writing distractors although it may not always write plausible distractors.

Table 8. *Distractor analysis.*

	Functions as intended	Selected by none	Selected by the upper group more	Selected equally by upper and lower group
Instructor	43 (71.6%)	6 (10%)	6 (10%)	5 (8.3%)
ChatGPT	34 (56.6%)	12 (20%)	1 (1.66%)	13 (21.6%)

3.3. The Results of Qualitative Data Analysis

Students were asked to share their perceptions of the question in two sets briefly. The answers were not rich in that sense. Although they "did not see a big difference between them," students had conflicting perceptions of the instructor's and ChatGPT's test items in some respects. One such perception is about the difficulty of the item sets. It seems that students related to the questions differently as some found the instructor's items more difficult while some others thought the opposite as evident from the following samples on the instructor's and ChatGPT's test items respectively.

This part was harder than the other one.

I think questions are same but difficulty of questions got higher in this section.

Another issue is with the clarity of the questions. Some students did not find the instructor's set clear while it was the opposite for some others. For example, one student said on the instructor's items:

There are some questions which are unclear. Seemingly there are two correct answers in one question.

Commenting on ChatGPT's items, on the other hand, students said the following:

Some of the questions were longer and were a bit harder to understand.

Questions are more complicated and confusing but the rest are easier. Questions are too long and also options. That's why it is confusing.

Conversely, for some other students, "the questions are great. They are easy to understand and clear." As indicated by the quotations above, the students found that ChatGPT's items were longer, which they believed made them more confusing and difficult. Yet, when the length of the stem and alternatives in the number of words are considered, the data does not support this perception as the average length of the stems in the instructor's set is 16.65 words, while it is 13.9 words for ChatGPT. When it comes to the length of the alternatives, however, ChatGPT wrote slightly longer alternatives as the length is 6.56 words for ChatGPT, whereas it is 6.12 words for the instructor. However, this difference is not large.

One other factor that some students highlighted was one related to vocabulary. For some students, ChatGPT's questions included more unfamiliar words, making the test more challenging as indicated by the following remark.

It was much more difficult than the other 20 questions. It wants more information. There are so many words I did not know.

Finally, students also had conflicting ideas about the aim of the questions written by the instructor and ChatGPT. For some of them, the instructor's questions asked for general information while ChatGPT asked for specific information. It was the opposite for some other students. For example, these are comments about the instructor's set.

The questions were mostly about our knowledge on the general information about the type of assessments.

These questions are mostly based on the course content, required students well-understanding of the course and requires knowledge of specific items.

The following comments, on the other hand, are on the questions written by ChatGPT.

The questions required some specific knowledge from us.

They are more flexible. Students might answer those questions with general understanding of course content and students can answer them with common sense.
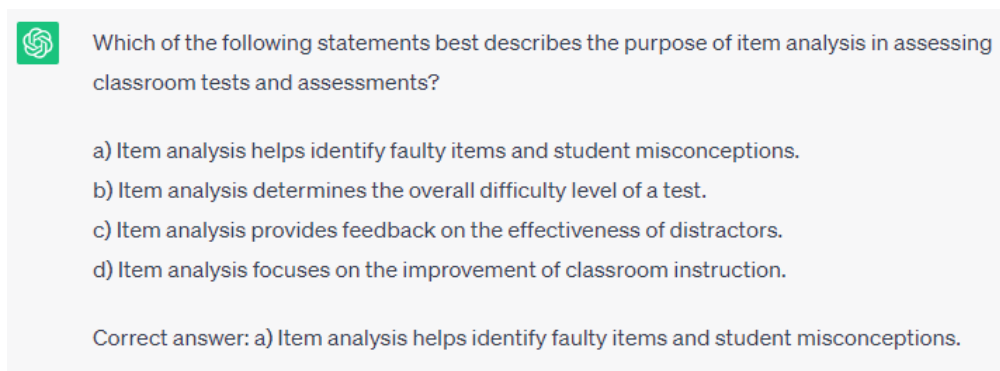
As the analysis of the students' comments demonstrates, they had conflicting views on various aspects of the questions written by the instructor and ChatGPT.

4. DISCUSSION

The results of the study demonstrate that ChatGPT can be utilized to create classroom tests that are on par with instructors' tests as this case study shows. The chatbot was able to create multiple-choice items and when they were put together and implemented, the test produced an acceptable degree of reliability. Since the items were created according to set learning outcomes and content, supposed content validity is likely to be high. However, there were also issues with the process.
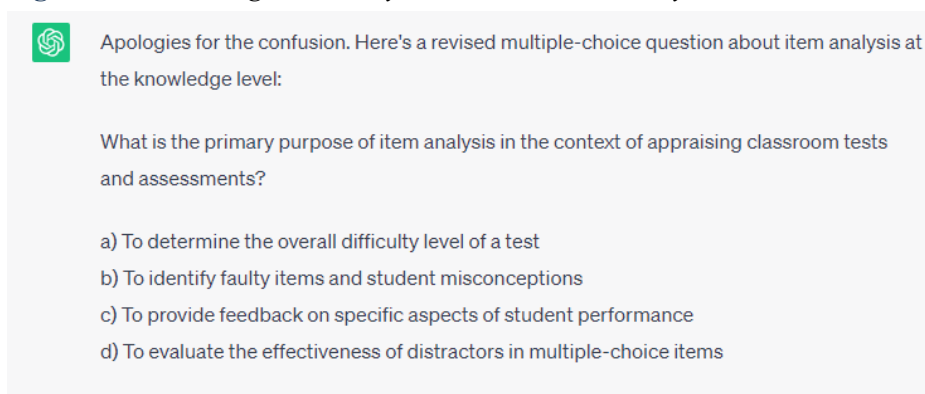
As mentioned in the methods section above, the test was created according to a table of specifications. To generate a test using ChatGPT to fit the table of specifications, I asked ChatGPT to create questions one by one by providing content. Yet, ChatGPT was not always apt to create items that were free of error. In other words, it created incorrect items. For example, when asked to create a test item about item analysis, it generated the question shown in Figure 1.

Figure 1. *An item generated by ChatGPT on item analysis*

Which of the following statements best describes the purpose of item analysis in assessing classroom tests and assessments?

a) Item analysis helps identify faulty items and student misconceptions.
b) Item analysis determines the overall difficulty level of a test.
c) Item analysis provides feedback on the effectiveness of distractors.
d) Item analysis focuses on the improvement of classroom instruction.

Correct answer: a) Item analysis helps identify faulty items and student misconceptions.

As can be seen, there is not one clear correct answer to this question, although ChatGPT shows a correct answer. Other options are also among the purposes of item analysis. When asked to regenerate by highlighting the problem, ChatGPT regenerated a similar question, still with the same problem of having more than one correct answer, as shown in Figure 2.

Figure 2. *An item regenerated by ChatGPT on item analysis*

Apologies for the confusion. Here's a revised multiple-choice question about item analysis at the knowledge level:

What is the primary purpose of item analysis in the context of appraising classroom tests and assessments?

a) To determine the overall difficulty level of a test
b) To identify faulty items and student misconceptions
c) To provide feedback on specific aspects of student performance
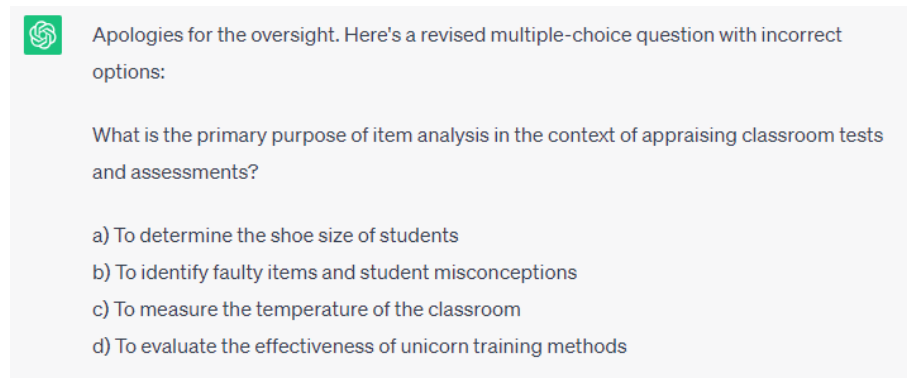d) To evaluate the effectiveness of distractors in multiple-choice items

When the issue of having more than one correct answer was fed into ChatGPT, it generated one with one outstanding correct answer. Yet, this time the other alternatives were nonsensical. As Figure 3 shows, all three items were utterly irrelevant. It generated, for example, an alternative that read "to evaluate the effectiveness of unicorn training methods."

This shows that ChatGPT is not always a reliable source to generate tests. It should be used with caution as it can pose a variety of failures (see, for example, Borji, 2023). Any test that ChatGPT generates should be closely scrutinized for any erroneous items. If specific items are obtained from ChatGPT, the instructor should also look for any possible errors like the ones
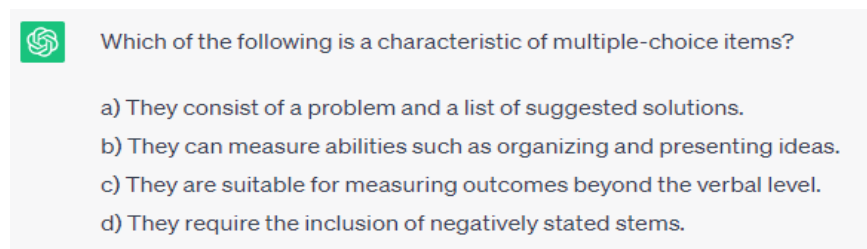
highlighted above. This leads to a legitimate question: does this take away some of the burden from the educators or rather add a more challenging task of evaluating items? This has to be carefully considered before making judgments about utilizing AI tools like ChatGPT for this purpose and in this manner.

Figure 3. *An item regenerated by ChatGPT on item analysis*

Apologies for the oversight. Here's a revised multiple-choice question with incorrect options:

What is the primary purpose of item analysis in the context of appraising classroom tests and assessments?

a) To determine the shoe size of students
b) To identify faulty items and student misconceptions
c) To measure the temperature of the classroom
d) To evaluate the effectiveness of unicorn training methods

Another issue was to get ChatGPT to generate questions according to the learning outcomes at different levels in the cognitive domain (Anderson et al., 2001; Bloom et al., 1956). To follow the table of specifications, the prompts indicated the cognitive level the question targets. For example, ChatGPT generated the following question to a prompt where asked to create a question at the level of understanding.
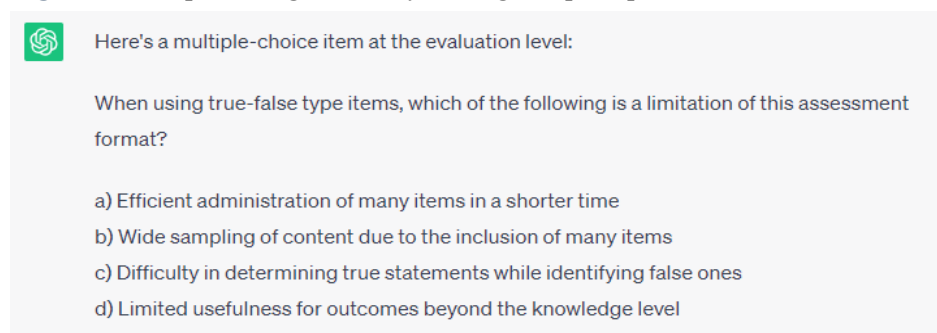
Figure 4. *The question generated for the given prompt*

Which of the following is a characteristic of multiple-choice items?

a) They consist of a problem and a list of suggested solutions.
b) They can measure abilities such as organizing and presenting ideas.
c) They are suitable for measuring outcomes beyond the verbal level.
d) They require the inclusion of negatively stated stems.

This question (Figure 4) is not at the understanding level. One factor that differentiates items at the level of understanding from those at the level of remembering is novelty (Miller et al., 2013). ChatGPT used the wording in the content provided to it to generate the question, which means that the students can memorize the content and simply answer this question without demonstrating their understanding of the content.

The following figure (Figure 5) points to the same problem. The chatbot was asked to create a question at the level of evaluation, yet ChatGPT failed to generate one at that level as it used the same content provided to it, producing a question at the level of remembering instead.

Figure 5. *The question generated for the given prompt*

Here's a multiple-choice item at the evaluation level:

When using true-false type items, which of the following is a limitation of this assessment format?

a) Efficient administration of many items in a shorter time
b) Wide sampling of content due to the inclusion of many items
c) Difficulty in determining true statements while identifying false ones
d) Limited usefulness for outcomes beyond the knowledge level

In a study by Moore et al. (2022), GPT-3 was used to evaluate the short-answer questions generated by the students. One thing that the researchers had GPT-3 do was to assign questions

to Bloom's taxonomy levels. Results indicate that GPT-3 failed to match the expert judgment of the cognitive level in 68% of the questions. GPT-3 also assigned 17 questions (14%) to evaluate and create levels that did not exist according to the pedagogical expert. The results of the cited study are relevant to the procedures followed in the current study where ChatGPT did not create questions at the intended level of Bloom's taxonomy. As these examples demonstrate, instructors should approach ChatGPT with caution. If the items generated by ChatGPT are used with confidence, then the tests created may not meet the need or may have low content validity, which can simply mean that they test "nothing" (Thompson & Vacha-Haase, 2018, p. 231).

In this case study, the analyses conducted have been post-hoc type such as reliability coefficient, item difficulty index, discrimination index, and distractor analyses. One potential issue with such post-hoc analyses is that the quality of the items is not tested prior to giving them to students and facing the risk of testing the students with low-quality items (Moore et al., 2023) and as such Moore et al. (2023) suggest a priori rule-based evaluation of items prior to using them for assessment. Still, both methods can be used in tandem to ensure the assessment of students' performance with the right tools and instruments. Even questions that are pre-evaluated can be analyzed through post-hoc techniques to ensure sound assessment, and it seems that both item generation and item evaluation can be handled with the assistance of AI tools such as ChatGPT before implementing specific classroom assessment tasks. Such AI tools are likely to be utilized for post-hoc analyses as well if the results are fed into them. Thus, AI tools can make assessment and evaluation tasks potentially less onerous for course instructors than they are now with the right content and prompts fed to them.

One interesting result in the reliability analysis in this case study is that the coefficient calculated for the combined test, including the instructor's and ChatGPT's items, was higher than the individual sets of tasks written by the instructor and ChatGPT alone. This finding is interesting as it may indicate that a combination of human and AI contributions may lead to an improved procedure. Halaweh (2023) asserts that "educators should encourage the use of human-AI tool augmentation for performing tasks such as finding information and ideas, editing texts and improving writing. By combining ChatGPT and human authors, the output is superior in terms of creativity, originality, and efficiency than if either one was to work alone" (p. 4).

As mentioned above, the items generated by ChatGPT were monitored by the instructor to establish that they fit with the table of specifications. Thus, the chatbot did not write a whole exam independently. This close monitoring of the questions may not reflect the independent use of AI language models to generate exams. This is relevant to González-Calatayud et al.'s (2021) contention that "this technology needs to be humanized. Research so far shows that a machine cannot assume the role of a teacher, and the way artificial intelligence works and carries out processes in the context of teaching is far from human intelligence" (p. 12). There may be ways to have AI tools to generate exams for the intended purposes of a class teacher, yet the experience in this study supports this position. The instructor needed to guide ChatGPT in preparing a test. Future research may reflect on comparing different exam generation methods like those including different degrees of contribution by the human or lack thereof.

One of the factors that are considered in addition to validity and reliability is practicality, or usability, which is related to factors such as economy, convenience, applicability, and the like (Miller et al., 2013; Thorndike & Thorndike-Christ, 2014). Since chatbots like ChatGPT or other similar AI tools can save time and effort on the part of the teachers if implemented efficiently, it would not be wrong to argue that they can increase the practicality of a test, and as such they can be said to potentially contribute to an important aspect of measurement and evaluation.

5. CONCLUSION

Research may indicate the utility of AI technologies like ChatGPT and may validate their effectiveness for use in education. However, there is also the practical aspect of the matter. Even when the use of such tools is strongly supported by empirical evidence in experimental conditions, how ready the teachers are for them is another essential issue. Wang et al. (2021) investigated, for example, the factors influencing teachers' intention to use AI technologies in teaching and found that perceived ease of use and self-efficacy were the most influential factors leading to teachers' behavioral intention to use AI technologies. They conclude that if action is taken to train teachers to enhance their self-efficacy beliefs, their attitude towards AI technologies and further intention to use them will likely increase. Thus, without incorporating such tools and their use into teacher training programs, informed practices about AI technologies to benefit teachers' experiences and students' learning will be a challenging task. Nazaretsky et al. (2022) share similar sentiments. In their research, the teachers may develop trust in AI technologies if they understand AI, AI-related technologies, and their usefulness and suggest professional development programs should include such components. In their study, teachers understood how AI works in assessing with a rubric and became more accepting of the procedures incorporating AI technologies in assessment. This indicated that to seriously consider incorporating AI tools in education, both teacher education programs and in-service training programs should be revised to include modules to prepare teachers for AI-supported practices. It is not only relevant at the individual level, organizations may also be AI-ready. Luckin et al. (2022), for example, propose a contextualized 7-step framework that will be tailored to the needs of the specific organization to help them with AI readiness.

This study aims to test the utility of ChatGPT in one aspect of the educational process in simulated testing rather than a real test situation where students would receive grades. This study was also limited to a compact group of learners enrolled in a single course at a university. More comprehensive studies eliminating such limitations are needed to further research on the issue.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Final International University, Ethics Committee, 2023/019/03.

Orcid

Mehmet Kanık https://orcid.org/0000-0002-1737-7678

REFERENCES

Al-Worafi, Y.M., Hermansyah, A., Goh, K.W., Ming, L.C. (2023). Artificial intelligence use in university: Should we ban ChatGPT? preprints.org, 2023020400. https://doi.org/10.20944/preprints202302.0400.v1

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for teaching, learning, and assessment: A revision of Bloom's taxonomy of educational objectives*. Longman.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The Classification of Educational Goals.* David McKay.

Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Chen, G., Yang, J., Hauff, C., & Houben, G.J. (2018). LearningQ: A large-scale dataset for educational question generation. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (Vol. 12, No. 1). Association for the Advancement of Artificial Intelligence.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 1-22.

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement*. Prentice-Hall.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?". *Journal of Computer Assisted Learning*, *37*(5), 1207-1216.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467.

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, *15*(2), ep421. https://doi.org/10.30935/cedtech/13036

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences. 13*(4), 410, 1-15. https://doi.org/10.3390/educsci13040410

Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence, 3*, 100076.

McCowan, R.J., & McCowan, S.C. (1999). *Item analysis for criterion-referenced tests*. Center for Development of Human Services. https://files.eric.ed.gov/fulltext/ED501716.pdf

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. https://doi.org/10.2139/ssrn.4354422

Miller, D.M., Linn, R.L., & Gronlund, N.E. (2013). *Measurement and assessment in teaching*. Pearson.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. In I. Hilliger, P. J. Munoz-Merino, T. D. Laet, A. Ortega-Arranz & T. Farrell (Eds.), *Educating for a new future: Making sense of technology-enhanced learning adoption* (pp. 243-257). Springer.

Moore, S., Nguyen, H.A., Chen, T., & Stamper, J. (2023). Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. In O. Viberg, I. Jivet, P. K. Munoz-Merino, M. Perifanou & T. Papathoma (Eds.), *Responsive and Sustainable Educational Features* (pp. 229-245). Springer.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British journal of educational technology, 53*(4), 914-931.

Okonkwo, C.W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033

Olney, A.M., Pavlik Jr, P.I., & Maass, J.K. (2017, June). Improving reading comprehension with automatically generated cloze item practice. In International Conference on Artificial Intelligence in Education (pp. 262-273). Cham: Springer International Publishing.

Reynolds, C.R., Livingston, R.B., & Willson, V. (2009). *Measurement and assessment in education.* Pearson.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., ... & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075.

Thompson, B., & Vacha-Haase, T. (2018). Reliability. In C. Secolsky and D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 231-251). Routledge.

Thorndike, R.M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education.* Pearson.

Van Dis, E.A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C.L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7

Wang, Y., Liu, C., & Tu, Y.F. (2021). Factors affecting the adoption of AI-based applications in higher education. *Educational Technology & Society, 24*(3), 116-129.

Yang, A.C.M., Chen, I.Y.L., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, *24*(3), 147–158. https://www.jstor.org/stable/27032862