Revolutionizing Assessment: Leveraging ChatGPT for Automated Item Generation: An AI Driven Exploratory Study with EFL Teachers

Ahmad A. Alsagoafi[1] & Hanan S. Alomran[2]

[1] Associate Professor, Department of English Language, College of Arts, King Faisal University, Alhasa, Saudi Arabia. E-mail: aalsagoafi@kfu.edu.sa

[2] Postgraduate Student, Department of English Language, College of Arts, King Faisal University, Alhasa, Saudi Arabia. E-mail: 223002326@student.kfu.edu.sa

Correspondence: Ahmad A. Alsagoafi, Associate Professor, Department of English Language, King Faisal University, Alhasa, Saudi Arabia. E-mail: aalsagoafi@kfu.edu.sa

Received: February 6, 2024 Accepted: April 23, 2025 Online Published: July 4, 2025

doi:10.5430/wjel.v15n6p385 URL: https://doi.org/10.5430/wjel.v15n6p385

Abstract

ChatGPT is gaining widespread acceptance in many disciplines since its launch at the end of 2022. The impact of ChatGPT on education is evident, but there is a dearth of knowledge on how English as a Foreign Language (EFL) teachers benefit from this technology. Therefore, this study investigates the use of ChatGPT to generate exam questions among EFL educators in Saudi Arabia. Through a mixed-methods approach that included an online questionnaire and an experimental design, the study attempted to gain insights from educators on using artificial intelligence (AI) technology for assessment. An online questionnaire was shared with 200 public school EFL teachers at various grade levels in the Eastern Province of Saudi Arabia. The findings revealed a varied landscape of perspectives, with some educators approving ChatGPT's efficiency in generating exam questions, whereas others expressed concerns about its limited application. A further examination of the instructor-designed and ChatGPT-generated test items revealed that ChatGPT has the potential to stimulate critical thinking and expand assessment formats. The results indicate that educators require professional development to leverage AI technology responsibly. Furthermore, this study highlights the importance of navigating the emerging ChatGPT in EFL classrooms to ensure reliability and consistency of the evaluation process.

Keywords: ChatGPT, assessment, mixed-methods approach, prompt design, EFL contexts

1. Introduction

ChatGPT, a cutting-edge language model that is powered by artificial intelligence (AI) and designed using OpenAI, is gaining widespread acceptance in many disciplines since its launch at the end of 2022. ChatGPT employs natural language processing to gather information from the internet and generate human-like responses to the inputs received from users. This AI technology is also trained to predict the next word in a sentence from a large text dataset. In the field of education, there are ongoing debates on the role of ChatGPT. While some educators see this AI technology as a way to support students' learning and reduce teachers' workload, others view it as a threat to education, as it could lead to cheating and plagiarism (Xie et al., 2023).

Despite the concerns about ChatGPT, there are calls to embrace its use in education for innovative learning and teaching purposes. This call for embracing AI technology is due to the expectations that ChatGPT will become a regular tool to assist with writing tasks in much the same way that calculators and computers have become integral to science and math (McMurtrie, 2022). Some scholars, like Sharples (2022), already view ChatGPT as an embedded component of education and think that its power should be harnessed to support student learning and development. Fortunately, the technological advances of ChatGPT are allowing for processes that can safeguard academic integrity. For example, Turnitin, a leading similarity detection company, is now using AI detection tools to distinguish between AI and human-written texts. According to Turnitin (2023), ChatGPT can be beneficial when used responsibly to enhance students' learning, but it is important to recognize novel challenges while taking advantage of the opportunities of this technology.

The Duolingo English Test (DET), an online language assessment tool, is another example of how ChatGPT can be used for generating test items. This computer-adaptive proficiency test can measure English proficiency among university candidates. The DET assesses the following four skills: speaking, writing, reading, and listening. The reading passages and multiple-choice questions for the DET are automatically generated using Generative Pre-trained Transformer 3 (GPT-3) (Park et al., 2022).

The impact of ChatGPT on education is evident, but there is a dearth of knowledge on how English as a Foreign Language (EFL) teachers benefit from this technology. To address this information gap, this study was conducted using an online questionnaire and an experimental design that involved generating questions by humans and comparing them to those generated by ChatGPT in a controlled environment. The study sought to answer the following research questions (RQs):

1) What are the perceptions of EFL teachers when using ChatGPT in assessments?
2) How do EFL teachers use ChatGPT to generate test questions?
3) To what extent does ChatGPT facilitate the assessment process for EFL teachers?

2. Literature Review

2.1 Language Assessment and AI

In the language assessment field, there are currently calls to forge closer ties with other neighboring fields such as cognitive and computational neuroscience, as well as AI (Aryadoust, 2023). The purpose of such an interdisciplinary approach is to stay abreast with the changes and requirements of discipline-specific assessment activities in the 21st century. The Internet disrupted mainstream education in the past, which is similar to what AI is presently doing. In particular, assessments in the field of English language gradually moved from traditional paper- and computer-based testing to Internet-based testing in the early 2000s. The rise of the Internet has made a lot of high-tech processes more manageable. For example, technology-mediated language assessment practices such as adaptive testing and automated assessment have become more prevalent (Saville & Buttery, 2023). Since the outbreak of COVID-19, technology-mediated language assessment tools have experienced radical changes and, as a result, an increase in use within the field of education. These unexpected changes are often referred to as the 'new normal.' Although online assessment materials were available before the pandemic, the integration of technology and assessments was not necessary. This is not to say that language testers were ignorant about the effect of technology on assessment. On the contrary, Winke and Isbell (2017) claimed that language assessments motivated by AI computer-assisted language assessments are "becoming normalized" (p. 313). Consequently, technology advancements "without which the language testing enterprise could have come to a complete halt or been severely damaged at best" (Sadeghi & Douglas, 2023, p. 2) have opened new avenues for language testing, such as at-home testing or remote assessment (Inoue et al., 2021)

However, the migration from paper- and computer-based tests to technology-integrated tests has given rise to concerns about test validity and reliability. The validity question needs to be reconceptualized so that the focus is not only on *what* is measured but also on *how* it is measured. This conceptualization will assist in understanding the impact of test methods on the performance of students. Therefore, it is crucial to thoroughly consider both theoretical and practical aspects of AI technology to prevent any potential negative consequences while ensuring that inferences about performance are valid and relevant to the construct being measured. This perennial concern about validity requires a fresh examination of the key issues involved in technology-mediated language assessment.

2.2 Automatic Item Generation for Assessment Purposes

AI-driven systems can generate test tasks such as adaptive testing questions, automated essay scoring, and language simulation scenarios. Specifically, the adoption of internet-based testing has created innovative solutions and opportunities for test task generation. AI innovations include item formats (Sireci & Zenisky, 2006), automated scoring (Shermis & Burstein, 2003), computer adaptive testing, and testing on-demand (Van der Linden & Glas, 2010; Wainer et al., 2000; Yan et al., 2014). Recent research on natural language processing with language modelling at its core has made significant contributions toward producing long and coherent texts by effectively predicting the next token in a sequence (Radford et al., 2019). This can benefit educational assessments by generating sample essays, providing high-quality practice materials, and automated essay scoring for objective feedback.

As such, trained AI models have the power to generate representations of language in the form of inputs such as text classification and question answering (Yang et al., 2019). In other words, models trained on large and diverse datasets are expected to exhibit promising performance in different datasets and domains (Radford et al., 2019). In the context of ChatGPT, Brown et al. (2020) discussed AI language models such as ChatGPT and claimed that "very large language models may be an important ingredient in the development of adaptable, general language systems" (p. 9). ChatGPT is an innovative language model that can mimic the format and style of natural text. For this reason, ChatGPT can be used to generate test tasks "by allowing test designers to prototype and iterate on new item types without the need for significant expert-annotated data or lengthy model development and training processes" (Attali et al., 2022, p. 2). In addition, Circi et al. (2023) identified four advantages of automatic item generation (AIG) for assessment development: time saving, cost efficient, rapid item development, and customized assessment and learning needs. Similarly, Pugh et al. (2016) suggested that AIG has the potential to assess higher-order skills and generate items with psychometric properties similar to traditional test formats. Despite these benefits, AIG is not widely used in educational assessments (Circi et al., 2023).

Circi et al. (2023) were able to identify three approaches to generate an item model: weak theory, cognitive/strong theory, and min-max. Among these, cognitive theory/strong theory was the most commonly used approach in education. This approach involves the following steps: (1) identify the skills and knowledge needed to solve the problem, (2) name the subject matter experts who will create the cognitive models, (3) develop item models based on the cognitive models, and (4) operate the item models using a computer-based algorithm. For example, Fridenfalk (2013) developed the Virtual Mathematics Assistant, an automatic item generator, to create formative assessments for the classroom. Similarly, Harrison et al. (2021) used an automated story generation via Markov Chain Monte Carlo sampling, which was trained to search in a diverse story corpus.

The interactive reading task in the DET is a recent innovative application of AIG. The test uses the GPT-3 language model to create reading passages along with comprehension questions, including correct answers and distractors (Attali et al., 2022). When given instructions for the desired text output (prompts), examples to guide the model in terms of format, subject matter, and style, and specific traits to refine the output, GPT-3 initially generated more than 14,000 reading passages. The genre ranged from news, expository, and

narrative texts. Attali et al. (2022) prompted GPT-3 with three to five examples that contained topics, titles, and passages about the genres. The reading passages were filtered by removing passages that contained a high degree of repetition to become 100–175 words with 5–20 sentences. After the first round of filtering, the number of texts was reduced from 1400 to just 800 texts. Then, test items were generated for these texts. Owing to GPT-3's inability to generate items and distractors for some passages, only 789 passages were retained after the filtering process. Attali et al. (2022) generated a vocabulary test task, a text completion task, two comprehension tasks, a main-idea task, and a possible title task, followed by a content and fairness review by expert reviewers, where 454 of the 789 passages were retained.

Another example of an internationally recognized English testing system that has embraced AI technology is Pearson's Test of English (PTE). Since 2009, PTE has been a pioneer in computer-based testing, but this tool has also relied heavily on people for grading and evaluating its higher-level (B2 level and above) real-life language ability tests, including critical speaking skills, such as philosophical discourse. However, with the advent of AI, these tests do not seem to require the same level of human involvement. AI-based tools can grade, evaluate, or analyze their tests (Pearson, 2012).

2.3 Computerized Adaptive Testing

Computerized adaptive testing is another pioneering application of AI in the field of testing. This application is used to develop advanced procedures and machine learning techniques to generate and control tests that can be adjusted to individual abilities. This inventive approach transforms the traditional testing process, offering numerous advantages in terms of accuracy and efficiency. However, this tool has some disadvantages, such as technical issues, high development and maintenance costs, and security concerns.

AI-generated tests have the potential to generate personalized assessments based on the specific needs of individual learners. By analyzing and adapting to a learner's unique strengths, weaknesses, and learning style using machine-learning techniques, these activities can be tailored to provide a customized learning experience (Kohnke, 2023). Personalized AI-generated tests can provide targeted examinations, allowing learners to demonstrate their understanding and progress in a manner that is relevant to their individual learning journey. Through adaptive feedback and remediation, these tests can offer customized support, help learners identify problem areas, and enhance learners' overall learning outcomes, which is considered an effective learning experience.

Teachers usually differ in their approaches to using AI for exams. One of the most intellectual ideas was to use ChatGPT as a collaborator to complete questions initiated by teachers, such as generating alternatives in multiple-choice questions. In a recent study by Skrabut (2023), ChatGPT was prompted to provide each question with five responses, one of which had to be the correct answer. Typically, teachers require time to brainstorm and devise distractors to design a test, but with AI, this activity is becoming less time-consuming. However, the validity of tests generated using ChatGPT can be uncertain because it relies heavily on teachers' experience and knowledge.

3. Methodology

This mixed-methods study employed an online questionnaire to investigate the use of ChatGPT by EFL teachers to generate test questions. In addition, data were obtained from an experiment that involved comparing AI-generated exam questions with teacher-created questions.

3.1 Research Sample

The sample for this study was public school EFL teachers in Saudi Arabia. An online questionnaire was distributed to 200 EFL teachers at various school levels (elementary, intermediate, and secondary) in the Eastern Province of Saudi Arabia. A total of 169 teachers completed the questionnaire, resulting in a response rate of 84.5%. Among the participants, 15.4% were female and 84.6% were male, from six different schools, with two representing each educational level. Of these teachers, 35.5% were teaching at the high school level, 31.4% were from the intermediate level, and 33.1% were from the elementary level. Many of the participants had considerable teaching experience: 26% had 11–15 years of experience, and 12.4% had over 20 years of experience in the field.

3.2 Instrument and Data Collection

The researchers developed an online questionnaire, the main data collection instrument, which was validated by experts. The questionnaire comprised two parts. The first part collected teachers' demographic information, such as sex, teaching experience, and current school level. The second part consisted of 13 items that examined teachers' frequency of using ChatGPT, purpose of use, effectiveness, and satisfaction levels. A five-point Likert scale was used to capture responses. The questionnaire was distributed to teachers via email and the schools' official WhatsApp groups. All participants were provided with detailed information about the purpose of the study as well as instructions for completing the questionnaire. Participation in the study was voluntary. Informed consent was obtained from all participants in writing prior to conducting the study. Second, an experiment was conducted to compare an exam designed by an expert teacher with one generated by ChatGPT regarding the same material.

3.3 Data Analysis

The collected data were analyzed using an open-source statistical software Jamovi (The Jamovi Project, 2024). Descriptive statistics, including frequencies, percentages, means, and standard deviations, were used to analyze the quantitative data. Cronbach's α scale was calculated to ensure reliability. Data were analyzed in accordance with the RQs. After analyzing the data, the findings were classified into three groups based on the RQs (Table 1), and the questionnaire items were assigned to each of them (4 items about RQ1, 3 items about RQ2, and 6 items about RQ3; see Appendix B and Table 4 for more details). The experimental data were processed against 12 characteristics of a good test, which were divided into 2 groups, (1) practical characteristics: usability, acceptability, adequacy, purpose, and comparability; (2)

technical characteristics: test items, objectivity, validity, reliability, discrimination, standardization (Bassey & Amanso, 2020).

4. Results and Discussion

The reliability of the scale used to measure these responses was high, with a Cronbach's α coefficient of 0.972.

Table 1. Distribution of Likert scale responses on RQs aspects

	Strongly agree (%)	Agree (%)	Neutral (%)	Disagree (%)	Strongly disagree (%)
RQ1 aspects	6.95	25.15	37.72	15.09	15.09
RQ2 aspects	7.89	34.32	28.8	12.82	16.17
RQ3 aspects	16.27	33.23	32.54	9.76	8.19

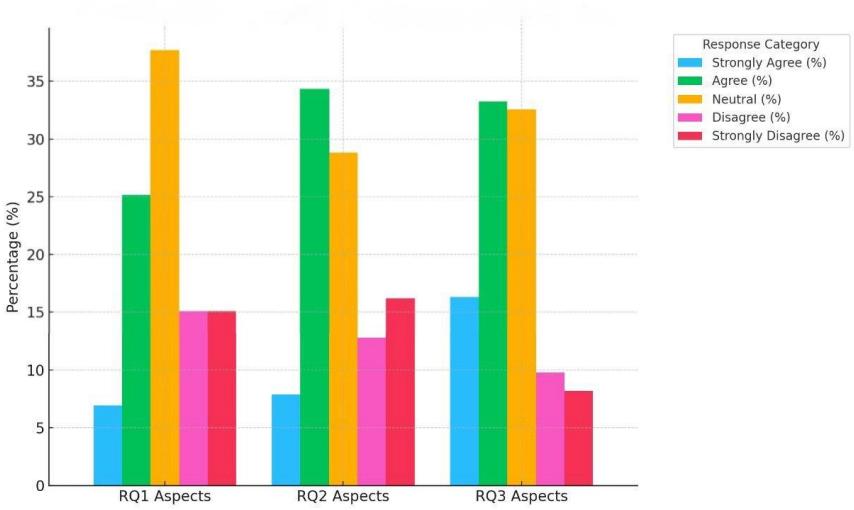

Figure 1. Distribution of Likert Scale Responses on RQ Aspects

4.1 RQ1: Perceptions of EFL Teachers Using ChatGPT

The distribution of responses showed that 32.1% (Strongly Agree + Agree) of teachers viewed ChatGPT positively for assessment purposes. However, a significant proportion of respondents, 37.7%, expressed neutrality, indicating mixed or uncertain perceptions. This hesitancy may reflect limited exposure to ChatGPT or insufficient confidence in its potential. This suggests that while some educators recognize the ChatGPT value, others might lack the hands-on experience necessary to form a definitive opinion. Disagreement (including Strongly Disagree) totaled 30.2%, highlighting a notable minority that is skeptical about the effectiveness or relevance of ChatGPT in classroom assessments. Such uncertainty may arise from concerns about ChatGPT's ability to align with specific pedagogical goals or its reliability in producing consistent results for diverse assessment contexts.

The findings revealed that the average mean score across responses was 3.06, demonstrating a general inclination among EFL teachers to actively encourage applying ChatGPT for language assessment purposes. This average aligns closely with the "Agree" category on the Likert scale, suggesting that many participants perceive ChatGPT positively in the matter of assessment. This positive perception aligns with the growing recognition of ChatGPT as a valuable aid in educational contexts, especially where automation can save time and streamline processes. The score reflects the overall acknowledgment of ChatGPT's effectiveness in generating test questions that align with its intended purpose. However, the placement of this score, which is slightly above the midpoint of the scale, suggests that ChatGPT's potential is recognized, but there are also reservations. Teachers may find ChatGPT effective for specific tasks but not necessarily a comprehensive solution for all assessment needs. This indicates that while ChatGPT may excel in certain areas, such as generating objective test questions, its limitations in handling more nuanced or subjective assessment tasks might contribute to the mixed perceptions. This nuanced response highlights the need for further exploration of why some educators are less confident or enthusiastic about the tool, perhaps owing to technical challenges, alignment with curriculum requirements, or a lack of familiarity with its functionalities. Addressing these reservations requires a deeper understanding of the contextual factors that influence teachers' confidence in adopting new technologies.

Furthermore, the standard deviation of 1.13 indicates moderate variability in the responses, which suggests that while there is a general trend toward agreement, there are notable differences in individual perceptions about applying ChatGPT in assessment. This variability could stem from factors such as differences in teaching experience, technological proficiency, and the extent to which teachers integrate ChatGPT into their workflow. For instance, teachers with higher technological proficiency may feel more optimistic about ChatGPT's capabilities, while those less familiar with digital tools might hesitate to rely on AI-driven solutions. Teachers with more experience using the tool may have developed strategies to harness its capabilities effectively, leading to stronger agreement with the positive statements. In contrast, those less familiar with or skeptical about AI technology might lean toward neutral or even negative responses, creating a wider

spread in the data. This variability emphasizes the importance of addressing individual differences through targeted professional development. By focusing on building teachers' confidence in AI tools and fostering a supportive environment for experimentation, these professional development efforts could bridge the gap between early adopters and hesitant users. Offering training sessions tailored to teachers' needs, including effective prompt design and best practices for integrating ChatGPT into assessments, could help reduce this disparity and foster more consistent positive perceptions across the teaching community.

In previous studies, researchers have raised concerns regarding the validity and reliability of AI tools, such as ChatGPT, in education (Aryadoust, 2023; Sadeghi & Douglas, 2023). The neutral and uncertain attitudes of teachers in this study appear to reflect similar worries, as many respondents did not seem to fully understand the potential of ChatGPT or were suspicious about its effectiveness in maintaining quality. This underlines the need to address these concerns through investigational evidence and case studies that demonstrate ChatGPT's practical benefits and limitations.

4.2 RQ2: How EFL Teachers Use ChatGPT to Generate Questions

Teachers showed stronger agreement with statements about their ChatGPT usage patterns, with 42.2% agreeing or strongly agreeing that they had used the tool effectively to generate test questions. Neutral responses (28.8%) from teachers suggest that some may have used ChatGPT sporadically or were still exploring its capabilities. This indicates that while there is a growing interest in using ChatGPT, some teachers may still be in the initial phases of experimentation, trying to determine how ChatGPT fits into their teaching practices. However, disagreements (28.9%) could be an indicator of technical challenges, lack of familiarity, or concerns about the tool's alignment with curricular goals. For example, teachers may find that ChatGPT-generated questions occasionally lack depth or fail to address specific skills required in language learning, leading to suspicion about its reliability.

The highest mean score of 3.23 indicates that EFL teachers strongly agree on the efficiency of using ChatGPT for generating test questions. This score, positioned closer to the "Agree" end of the Likert scale, reflects a positive perception of ChatGPT's role in assisting with the creation of assessment materials. This suggests that ChatGPT is seen as a practical and resourceful tool, especially by teachers seeking to save time and enhance creativity in test design. This finding suggests that many teachers find ChatGPT particularly helpful for generating test questions efficiently, saving time, or offering creative alternatives to traditional question design methods. Moreover, the efficiency of ChatGPT may help reduce the cognitive load involved in designing test items, allowing teachers to focus more on other pedagogical tasks. Such agreement might also point to ChatGPT's perceived ease of use and its ability to provide outputs that align well with the immediate requirements of EFL assessments. Teachers who were more comfortable using AI tools or had clearer objectives when employing ChatGPT likely contributed to the higher mean score, suggesting that when the tool is applied purposefully, it can yield favorable outcomes. This highlights the importance of encouraging AI tools in the field, as teachers who approach ChatGPT with specific goals tend to achieve more effective results.

The overall variability, reflected in a standard deviation of 1.15, highlights the diversity of how teachers use ChatGPT and perceive its effectiveness. This slightly higher variability suggests that not all teachers shared the same level of confidence in, or satisfaction with the tool. Such variability may reflect differences in teachers' levels of technological proficiency, as well as varying expectations regarding the quality of ChatGPT-generated questions. While some educators may see ChatGPT as a reliable partner in test creation, others may struggle with its application owing to factors such as unfamiliarity with crafting precise prompts, limitations in aligning AI-generated questions with specific learning outcomes, and concerns regarding the quality and appropriateness of the generated content. For instance, some teachers may feel that the lack of contextual understanding in ChatGPT-generated questions reduces their relevance to specific classroom goals, thus limiting their usability. This diversity in usage patterns underscores the varying levels of ChatGPT integration into teaching practices. For instance, some teachers may use it exclusively for inspiration or as a supplementary tool, while others may rely on it extensively for designing the entire assessment. This range of usage reflects the flexibility of ChatGPT, but also emphasizes the need for clear guidelines to ensure its effective adoption. This variability indicates the need for more structured guidance on how ChatGPT can be used effectively in diverse educational contexts, ensuring that teachers across different levels of proficiency and comfort with AI tools can benefit equally from its capabilities.

This finding aligns with the results of Attali et al. (2022) and Circi et al. (2023), who showed that ChatGPT can be used for creating and customizing test items. Both studies underscored the adaptability of ChatGPT in tailoring test content to meet specific learning objectives, which echoes with the responses from teachers in this study. Furthermore, consistent with the findings of these studies, participants in this study also reported that saving time and generating creative test questions were advantages of AI technology. The ability to quickly generate a variety of question types, including multiple-choice, short answer, and matching questions, was particularly valued by teachers. According to Skrabut (2023), AI technology can also be used to generate distractors and alternative questions. In this study, teachers who professionally used ChatGPT were more likely to achieve better results in generating high-quality test questions, emphasizing the importance of training in the design and application of AI technology. This suggests that effective training in using ChatGPT, particularly in constructing purposeful prompts, can significantly enhance the quality of outputs and the overall user experience.

4.3 RQ 3: ChatGPT's Facilitation of the Assessment Process

Teachers' responses were varied regarding ChatGPT's ability to facilitate the assessment process. While 49.5% of teachers agreed or strongly agreed that ChatGPT could facilitate assessment processes, a significant 32.5% remained neutral. The neutral responses suggest that a considerable number of teachers are still unsure about the full extent of ChatGPT's capabilities, possibly due to limited practical

experience or a lack of confidence in its outputs. Interestingly, disagreement was lower, at 17.9%, suggesting that only a few teachers actively found ChatGPT unhelpful. This finding indicates an overall acknowledgment of ChatGPT's effectiveness in simplifying some aspects of assessment, such as question generation, but also highlights its inability to handle all the complexities of assessment design. For example, while many educators might appreciate the speed and efficiency of AI-generated content, they may also recognize that these outputs require further adaptation to meet broader pedagogical requirements.

The category with the lowest mean score of 2.60 reflects more neutral or slightly negative perceptions of ChatGPT's effectiveness in simplifying the assessment process for EFL teachers. Positioned closer to the "Neutral" range of the Likert scale, this score suggests a sense of hesitation or uncertainty among teachers regarding the tool's ability to streamline assessment tasks comprehensively. This highlights that, while ChatGPT may perform well in certain areas, it has not yet achieved the level of sophistication required to address the multifaceted nature of educational assessments. This finding could indicate that, while ChatGPT is recognized for its efficacy in generating test questions, it may fall short of addressing the broader complexities of assessment design, such as aligning questions with specific curriculum objectives, ensuring fairness, and accommodating diverse student needs. These limitations may stem from the fact that ChatGPT lacks the contextual understanding and critical judgment that human educators bring to the assessment process, particularly when tailoring materials to meet the unique needs of a class or individual students. Additionally, teachers might perceive certain aspects of assessment, such as interpreting student performance or ensuring question validity, as tasks that require a level of human judgment and expertise that ChatGPT cannot fully replicate. This suggests that while ChatGPT can function as a helpful tool, it is not currently equipped to replace essential human involvement in assessment design. The lower mean score underscores that, while ChatGPT has potential, its application in the assessment process is perceived by many as limited or supplementary rather than transformative.

Interestingly, the responses for this category displayed less variation, with a standard deviation of 1.11. This statistic suggests that teachers' perceptions of ChatGPT's role in simplifying assessments were relatively consistent, with fewer extreme opinions compared with other aspects of its use. The consistency in responses may indicate a shared understanding among educators of both the tool's strengths and its limitations, particularly in the context of creating assessments for EFL instruction. This uniformity might indicate that the tool's limitations in facilitating a broader assessment process are universally recognized, regardless of individual differences in experience or familiarity with ChatGPT. Teachers may agree that while ChatGPT can be a valuable aid, it requires significant human input to ensure that its outputs meet educational standards and learning objectives. The lack of strong variability also suggests that ChatGPT is not yet fully equipped to handle all aspects of educational assessment in the EFL context. To address this issue, it may be beneficial to provide additional tools or resources that complement the capabilities of ChatGPT, such as frameworks for adapting its output or integrating its use with human oversight. For instance, offering guidelines on how to enhance AI-generated assessments or blend ChatGPT with other AI tools might help address its current limitations and expand its efficacy in varied educational contexts. By acknowledging its current limitations while leveraging its strengths, educators can employ ChatGPT as part of a balanced approach to assessment design.

In these findings, teachers have shown a pattern of perceptions about ChatGPT's capabilities, viewing it as a supplementary tool rather than a complete solution for assessments, particularly due to a lack of knowledge about the features of AI technology. This highlights the need for targeted training programs aimed at increasing teachers' familiarity with AI tools, ensuring they can fully operate ChatGPT's potential while understanding its limitations. Previous studies have pointed to the same issue, considering this limitation (see Attali et al., 2022; Circi et al., 2023; Pugh et al., 2016; Sadeghi & Douglas, 2023). These studies emphasize that while ChatGPT can enhance efficiency and creativity in assessment design, it requires human expertise to adapt outputs to specific educational contexts.

4.4 Practical Experiment

RQ2, which asked EFL teachers about the use of ChatGPT to generate test questions, aroused curiosity regarding the disparities between the questions generated by ChatGPT and those prepared by teachers. To explore these differences and ChatGPT's ability to create exam questions, an experiment was conducted to compare a senior English teacher's test with a test generated by ChatGPT. Both were given the same text (Appendix A: Table 1) and asked to generate five multiple-choice questions, each with four response options (labeled A, B, C, and D) (see Appendix A, Table 2 and Table 3). The questions were then categorized according to various aspects, as shown in Table 5.

Table 5. Comparison between teacher-made tests and ChatGPT-generated tests

Aspect	Teacher version	ChatGPT-generated version
Skills to be tested	Comprehension, vocabulary, identifying facts, and details	Comprehension, inference, critical thinking, vocabulary
Question structure	Simple	Complex
Difficulty level	Quite easy	Reasonable
Focus of questions	Direct information	Analyzing and interpreting information
Evaluation skills	Basic understanding and recall	Inference, reasoning, and interpretation
Depth	Narrow	Deep
Use of relative information	Slight	Wide
Emphasis on critical thinking	Limited	Noticeable
Quality of questions	Clear and concise, but lacks complexity	Well-structured and requires some thinking
Main points covered	Covers basic information and facts	Explores various aspects, including reasons, numbers, and threats

Based on an in-depth analysis of the two versions, the teacher version focused more on basic skills and information, whereas the ChatGPT-generated version placed greater emphasis on critical thinking and interpretation. Questions in the ChatGPT-generated version were more challenging and required a deeper understanding of the text.

5. Conclusion

In conclusion, this study was able to achieve several objectives that contributed to a deeper understanding of item generation using ChatGPT. According to this study, EFL teachers have mixed feelings regarding the use of ChatGPT to generate test questions. While most participants agreed that ChatGPT enhanced their assessment practices through question generation, others were skeptical of its ability to provide comprehensive solutions for all assessment needs. Participants familiar with ChatGPT and its prompt design were more optimistic about its effectiveness in generating formative assessments for the classroom. However, the consensus was that ChatGPT item generation is not yet fully capable of handling all aspects of educational assessment in an EFL context.

Even so, the results of this study should be interpreted with caution. Although the current study met the minimum requirements to run the analysis, the sample size was not sufficiently large to be generalized to the Saudi population. Moreover, this study was conducted specifically within the context of English language teaching in Saudi Arabia. Thus, the results cannot be generalized to other contexts or educational settings. Furthermore, this study only used a questionnaire and an experiment for data collection. For a more comprehensive and accurate picture of EFL teachers' use of ChatGPT to generate items, future studies should have a larger and more representative sample of participants. Likewise, future research should include other participants such as assessment experts, AI Specialists, or curriculum developers so that multiple sources of evidence can be considered.

Accordingly, there are some important recommendations to be considered by EFL teachers when using ChatGPT to generate questions. First, ChatGPT can be effectively guided to generate questions for testing specific language skills. Second, the more details the user includes in the prompts, the better the AI technology is at providing results. Finally, it is important to adjust the difficulty and complexity of the questions, either by selecting an appropriate complexity level or by specifying the type of learners for ChatGPT.

Finally, this study has pedagogical implications for language teaching and assessment. The results of this study showed that familiarity with ChatGPT and prompt design can enhance the performance of the tool. Therefore, professional development on how to harness ChatGPT to craft appropriate prompts is highly recommended for novice users. Additionally, this study pointed out that item generation through ChatGPT is not yet fully equipped to handle all aspects of educational assessments in the EFL context. To overcome this limitation, educational institutions could incorporate additional tools or resources that complement ChatGPT's capabilities, along with human oversight.

Acknowledgments

The authors extend their sincere appreciation to the participants of this study for their valuable contributions. The authors also gratefully acknowledge the support of the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, which made this work possible.

Authors' contributions

Dr. Ahmad Alsagoafi is the first and corresponding author of this paper. He came up with the idea after identifying the research gap and then thought of publishing it with his MA student, Miss Hanan Alomran. Both authors were responsible for the design and the revision of the study. Miss Hanan took charge of data collection, analysis, and discussion. Dr. Ahmad provided help and support in those parts and then worked together with her on the remaining sections, contributing heavily to the abstract, introduction, literature review, conclusion, and references. As the corresponding author, he secured funding for the study. All authors read and approved the final manuscript.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU251958].

Competing interests

Sample: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing*, *40*(1), 8-14. https://doi.org/10.1177/02655322221125204

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077. https://doi.org/10.3389/frai.2022.903077

Bassey, B. A., & Amanso, E. O. I. (2020). Assessing the technical and practical qualities of a good test as a measuring instrument. Prestige. *Journal of Counseling Psychology, 3*(2), 89-99. Retrieved from https://openaccessglobal.com/wp-content/uploads/2021/07/technical_and_practical_qualities_of_a_good_test.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). *Language models are few-shot learners*. arXiv. Retrieved from https://arxiv.org/abs/2005.14165

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

Fridenfalk, M. (2013). System for automatic generation of examination papers in discrete mathematics. *In Proceedings of the International Conference on Cognition and Exploratory Learning in Digital Age (CELDA)* (pp. 365–368). International Association for Development of the Information Society. Retrieved from https://files.eric.ed.gov/fulltext/ED562287.pdf

Harrison, B., Purdy, C., & Riedl, M. (2021). Toward automated story generation with Markov chain Monte Carlo methods and deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *13*(2), 191-197. https://doi.org/10.1609/aiide.v13i2.13003

Inoue, C., Khabbazbashi, N., Lam, D., & Nakatsuhara, F. (2021). Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices. (*IELTS Research Reports Online Series, No. 2*). British Council, Cambridge Assessment English, and IDP: IELTS Australia. Retrieved from https://www.ielts.org/teaching-and-research/research-reports

Kohnke, L. (2023). Microlearning with chatbots. In *Using Technology to Design ESL/EFL Microlearning Activities* (pp. 71–79). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-2774-6_7

McMurtrie, B. (2022, December 13). AI and the future of undergraduate writing. *The Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing.

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Duolingo English Test: Interactive reading* (Duolingo Research Report No. DRR-22-02). Duolingo. https://doi.org/10.46999/RCXB1889

Pearson. (2012). *The Official Guide to PTE Academic*. Essex: Pearson PTE.

Pugh, D., De Champlain, A., 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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* [Technical report]. OpenAI. Retrieved from https://www.semanticscholar.org/paper/9405cc0d6169988371b2755e573cc28650d14dfe

Sadeghi, K., & Douglas, D. (2023). *Fundamental considerations in technology mediated language assessment*. London: Routledge. https://doi.org/10.4324/9781003292395

Saville, N., & Buttery, P. (2023). *Interdisciplinary collaborations for the future of learning-oriented assessment*. London: Routledge. https://doi.org/10.4324/9781003292395-17

Sharples, M. (2022). Automated essay writing: An AIED opinion. *International Journal of Artificial Intelligence in Education*, *32*(4), 1119-1126. https://doi.org/10.1007/s40593-022-00300-7

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring*. New York: Routledge. https://doi.org/10.4324/9781410606860

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation.

In *S. M. Downing & T. M. Haladyna* (Eds.), *Handbook of test development* (pp. 343–362). Routledge. https://doi.org/10.4324/9780203874776-21

Skrabut, S. (2023). *80 ways to use ChatGPT in the classroom. Tubarks – The musings of Stan Skrabut.* Retrieved from https://tubarksblog.com/product/80-ways-to-use-chatgpt-in-the-classroom/

The Jamovi Project. (2024). *Jamovi*. (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org

Turnitin. (2023). *Plagiarism detector: Prevent academic misconduct*. Retrieved from https://www.turnitin.com/

Van Der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. https://doi.org/10.1007/978-0-387-85461-8

Wainer, H., Dorans, J., N., Flaugher, R., Green, F., B., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. New York: Routledge. https://doi.org/10.4324/9781410605931

Winke, P. M., & Isbell, D. R. (2017). Computer-Assisted Language Assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 313–325). Springer. https://doi.org/10.1007/978-3-319- 02237-6_25

Xie, Y., Wu, S., & Chakravarty, S. (2023). AI meets AI: Artificial intelligence and academic integrity: A survey on mitigating AI-assisted cheating in computing education. In The *24th Annual Conference on Information Technology Education (SIGITE '23), October 11–14, 2023, Marietta, GA, USA*. ACM. https://doi.org/10.1145/3585059.3611449

Yan, D., Von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. New York: CRC Press. https://doi.org/10.1201/b16858

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., & Lin, J. (2019). End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 72–77). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4013

Appendix A

Tigers

Who would win in a fight, a lion or a tiger? Well, if size has anything to do with the matter, the tiger would win. That's because tigers are the largest of all cat species. They grow up to 11 feet and weigh as much as 670 lbs. This makes tigers the third largest land carnivore. The only land carnivores that are larger are polar bears and brown bears. Tigers are not only large but also fast. They can sprint as fast as 40 miles per hour for short distances and leap as far as 30 feet horizontally. This makes for an extremely dangerous pounce. You might not think that such large, fast, and ferocious creatures need help to survive, but they do. The tiger is an endangered species.

Despite all of the tiger's strengths, the future of the species is uncertain. Tigers face a very high risk of extinction. It is estimated that at the start of the 20th century, there were over 100,000 tigers living in the wild. By the turn of the century, the number of tigers outside of captivity dwindled to just over 3,000. Interestingly, the most serious threats that tigers face come from a much smaller species, one with an average weight of around 140 lbs. That species is Homo sapiens, better known as humans. Humans threaten tigers in primarily two ways: hunting and destroying habitat.

Figure 2. Sample Text

1) Tigers are the largest _____ species.
   A) dogs B) cats C) camels D) fish
2) Homo sapiens means _____.
   A) humans B) tigers C) dinosaurs D) mummies
3) The top speed of tigers is_____.
   A) 10 MPH B) 20 MPH C) 100 MPH D) 40 MPH
4) Tigers could grow to _____ feet long.
   A) 12 B) 60 C) 11 D) 10
5) Tigers nowadays are facing a very high _____.
   A) risk B) number C) place D) rice

Figure 3. Teacher's version test

1) What is the main reason suggested in the text for why a tiger might win in a fight against a lion?

A) Speed B) Ferocity C) Size D) Agility

2) According to the text, what is the maximum weight a tiger can reach?

A) 670 lbs B) 800 lbs C) 500 lbs D) 1,000 lbs

3) What is the estimated number of tigers living in the wild at the start of the 20th century?

A) Over 100,000 B) 50,000 C) 10,000 D) 3,000

4) Which two bears are mentioned as larger land carnivores than tigers?

A) Polar bears and grizzly bears B) Brown bears and black bears

C) Grizzly bears and black bears D) Polar bears and brown bears

5) What are the two main threats to tigers mentioned in the text?

A) Pollution and climate change B) Poaching and disease

C) Hunting and habitat destruction D) Deforestation and natural disasters

Figure 4. ChatGPT's version test

Appendix B

Table 3. Statements from online questionnaire

Statements		Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Mean	Std. deviation	Rank
I feel shy to use ChatGPT to generate an exam for my students.	N	15	29	72	30	23	3.10	1.12	3
	%	8.9%	17.2%	42.6%	17.8%	13.6%			
I describe myself as a professional ChatGPT user.	N	11	41	67	23	27	3.08	1.13	4
	%	6.5%	24.3%	39.6%	13.6%	16%			
I know how to insert ChatGPT commands (prompts) properly.	N	14	47	54	25	29	3.05	1.20	5
	%	8.3%	27.8%	32%	14.8%	17.2%			
I always get what I am looking for exactly when I use ChatGPT.	N	7	53	62	24	23	3.02	1.08	6
	%	4.1%	31.4%	36.7%	14.2%	13.6%			
I always use prompts in my L1 to generate questions in L2.	N	5	49	50	32	33	3.23	1.15	1
	%	3%	29%	29.6%	18.9%	19.5%			
I tried ChatGPT previously to generate exam questions for my students.	N	13	52	44	23	37	3.11	1.27	2
	%	7.7%	30.8%	26%	13.6%	21.9%			
I do some editing and modification on AI generated exams.	N	22	73	52	10	12	2.51	1.03	11
	%	13%	43.2%	30.8%	5.9%	7.1%			
I think ChatGPT could generate questions for limited language skills.	N	10	57	72	16	14	2.80	0.984	7
	%	5.9%	33.7%	42.6%	9.5%	8.3%			
I am afraid that ChatGPT may limit my creativity in creating exams.	N	22	52	54	26	15	2.76	1.14	8
	%	13%	30.8%	32%	15.4%	8.9%			
I feel worried about the fairness of exam questions generated by ChatGPT.	N	23	54	62	19	11	2.65	1.06	9
	%	13.6%	32%	36.7%	11.2%	6.5%			
My main reason for using ChatGPT in assessment is its efficiency in meeting tight deadlines.	N	27	60	56	11	15	2.57	1.11	10
	%	16%	35.5%	33.1%	6.5%	8.9%			
I believe AI questions generated are good for quick quizzes only.	N	31	63	49	13	13	2.49	1.11	12
	%	18.3%	37.3%	29%	7.7%	7.7%			
I am concerned about the validity of the questions or tests generated by ChatGPT.	N	52	51	37	14	15	2.34	1.24	13
	%	30.8%	30.2%	21.9%	8.3%	8.9%			
Reliability							Cronbach's α = 0.972		