كفاءة استخدام شات جي بي تي في تصميم اختبارات فهم المقروء لطلاب اللغة الإنجليزية كلغة أجنبية

داليا عمر احمد
جامعة حلبجة – كلية التربية الأساسية،
قسم اللغة الإنجليزية، طالب ماجستير
dalya.ahmad@uoh.edu.iq

أ.م.د. نيان كامل غفور
جامعة حلبجة
nian.ghafour@uoh.edu.iq

الكلمات المفتاحية: الذكاء الاصطناعي، نماذج لغوية كبيرة، كفاءة، شات جي بي تي، كفاءة، كفاءة، اختبار مهارات القراءة.

كيفية اقتباس البحث

احمد ، داليا عمر، نيان كامل غفور ، كفاءة استخدام شات جي بي تي في تصميم اختبارات فهم المقروء لطلاب اللغة الإنجليزية كلغة أجنبية،مجلة مركز بابل للدراسات الانسانية، حزيران 2026،المجلد:16 ،العدد:6.

مفهرسة في Indexed
IASJ

مسجلة في Registered
ROAD

The Efficiency of Using ChatGPT in Designing EFL Reading Comprehension Tests

Dalya Omer Ahmad
University of Halabja-College Basic Education, English Department, M.A student

Asst. Prof. Dr.Nyan Kamil Ghafour
University of Halabja

Keywords : Artificial intelligence, Large Language models, ChatGPT, Efficiency, and Reading skill test.

How To Cite This Article
Ahmad, Dalya Omer , Nyan Kamil Ghafour , The Efficiency of Using ChatGPT in Designing EFL Reading Comprehension Tests,Journal Of Babylon Center For Humanities Studies, june 2026,Volume:16, Issue6.

الملخص

يعد تصميم اختبارات اللغة يدويا وضمان جودتها عملية معقدة تتطلب معلما خبيرا ذا خبرة، لا سيما في اختبارات فهم المقروء. يفتح الذكاء الاصطناعي افاقا جديدة لدعم تعليم اللغة عموما، وعملية التقييم خصوصا. تقارن هذه الدراسة بين اختبارات فهم المقروء التي صممها معلمو اللغة الإنجليزية كلغة أجنبية وتلك التي صممها برنامجChatGPT ، وذلك من خلال فحص أداء 51 طالبا جامعيا في السنة الثانية. تتضمن الاختبارات أسئلة في القواعد والمفردات ومهارات فهم المقروء. اجريت المقارنة بتطبيق كلا الاختبارين على نفس المجموعة من الطلاب في نفس الوقت والمكان. تم تحليل البيانات باستخدام برنامجSPSS ؛ حيث أشارت الإحصاءات الوصفية والتحليل متعدد المتغيرات باستخدام تحليل Pillai's Trace إلى وجود فروق دالة إحصائيا في أداء الطلاب بين الاختبارين وفي مهارات القراءة (المفردات والقواعد وفهم المقروء). بالإضافة إلى ذلك، أثرت أنواع الاختبارات والمهارات المستخدمة على أداء الطلاب. أظهرت

النتائج أن الاختبار المنشأ بواسطة ChatGPT أكثر فعالية وكفاءة في تقييم المفردات وفهم المقروء، بينما يعد الاختبار الذي يصممه المعلم أكثر فعالية في تقييم أداء الطلاب في قواعد اللغة. وعند تصميم الاختبارات المنشأة بواسطة الذكاء الاصطناعي ومراجعتها بشكل صحيح، يمكنها أن تضاهي أداء الاختبارات التي يصممها المعلم في تقييم مهارات القراءة. وتخلص الدراسة إلى أن ChatGPT يتمتع بقدرة عالية على مساعدة المعلمين في تصميم أسئلة اختبار موثوقة وصالحة. كما تبرز الدراسة أهمية اتباع نهج تكاملي يجمع بين الاختبارات المصممة بواسطة الذكاء الاصطناعي والاختبارات التي يصممها المعلم لتعزيز فعالية وكفاءة عملية تصميم اختبارات اللغة الإنجليزية كلغة أجنبي

Abstract

Designing language tests manually and ensuring their quality is a complex process that needs a knowledgeable teacher with experience, especially for reading comprehension tests. Artificial Intelligence (AI) opens up new ways to assist in language education, generally, and the assessment process, particularly. This study compares EFL teacher-made and ChatGPT-designed reading comprehension tests by examining the performance of 51 second-year university students on the tests in order to examine the efficiency of using ChatGPT compared to humans in designing language tests. The tests consist of grammar, vocabulary, and reading comprehension skill items, and each test is on 20 marks. The comparison was done by giving both tests to the same group of students at the same time and place. The data was analyzed through the use of SPSS Software; both descriptive statistics and multivariate analysis using Pillai's Trace for analyzing students' responses indicated that there are significant differences in students' performance across both tests and across the reading skills (vocabulary, grammar, and reading comprehension). Additionally, the test types and the skills affected students' performance. The findings showed that the test created by ChatGPT is more effective and efficient in vocabulary and reading comprehension, but the teacher-made test is more effective in assessing students' performance in grammar. When AI-generated tests are designed and reviewed properly, they can perform comparably to teacher-made tests in assessing reading skills. The study concludes that ChatGPT has a strong capability to assist teachers in designing reliable and valid test questions. The study highlights the value of a complementary approach that integrates AI-designed with teacher-designed tests to enhance the effectiveness and efficiency of the EFL language test design process.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

1. Introduction

The rapid advancement of technology has deeply transformed all aspects of education, including the assessment process and test design process in the way tests are designed, developed, and delivered (Jurāne-Brēmane, 2023; Sembey et al., 2024). Designing high-quality language tests, especially reading comprehension skills, will be highly affected by the question types and items, and requires lots of time, effort, and expert resources (Säuberli & Clematide, 2024; Sugawara et al., 2022). According to Ridwan (2024) and Verhoeven and Perfetti (2008), the design of reading comprehension tests not only depends on the design of authentic or real-world texts, but also on the test being designed in a way that can evaluate different areas of comprehension and understanding, such as grammar, vocabulary, and interpretation.

With the development of Artificial Intelligence (AI) technologies, especially in the field of Natural Language Processing (NLP), the interest in using these AI technologies for the assessment process development has increased to enhance and make the process easier and more successful (Victoria & Davier, 2023). Among those innovations of AI technologies, one of the chatbots that was greatly developed and focused on in the education process is large-scale transformer-based language models like OpenAI's GPT (Generative Pre-trained Transformer), which offers the potential for completely automated, human-like test item generation (Brown et al., 2020; Jeon & Lee, 2023; Singha et al., 2024). The chatbot models the free version and the paid version which generate texts and questions for designing tests with little or no human intervention by utilizing large pre-trained datasets and few-shot learning (Brown et al., 2020; Hansen & Hebart, 2022).

Since the release of ChatGPT, the application of AI in language learning and teaching has expanded rapidly (Kohnke et al., 2023; Shin, 2023). Because ChatGPT can produce diverse, cohesive, and contextually relevant text, which teachers can investigate and explore how it might be used to create reading passages, comprehension questions, vocabulary drills, and grammar assignments for students learning English as a foreign language (EFL) (Hassan & Alsalwah, 2025; Moon et al., 2025). This rapid development of AI-assisted test design technologies offers solutions to several challenges in educational assessment, including the need for continuous test item generation, item bank diversification, and reduced workload for teachers who are often constrained by time, resources, and repetitive item development cycles (Aryadoust et al., 2024; Gehringer, 2004; Kurdi et al., 2019).

This study wants to answer the research questions which are:

1. Is there any significant difference between the efficiency of the ChatGPT-designed language reading comprehension test and the teacher-made test across grammar?
2. Is there any significant difference between the efficiency of the ChatGPT-designed language reading comprehension test and the teacher-made test across vocabulary?
3. Is there any significant difference between the efficiency of the ChatGPT-designed language reading comprehension test and the teacher-made test across reading skills?

Additionally, this study aims to examine the use of the AI tool ChatGPT in generating EFL reading comprehension tests at the university level to assess the efficiency of using ChatGPT in the language test design process through making a comparison between students' performances on both tests. Also, comparing the students' scores on grammar, vocabulary, and comprehension skills to find whether ChatGPT, as an AI tool, is capable of designing an effective test item for these skills of reading.

2. Literature Review

This section starts by discussing relevant theoretical frameworks for the current study. Then it connects core assessment principles to practical test design to see if AI-designed EFL reading comprehension tests can be comparable to or superior to teacher-made tests in terms of grammar, vocabulary, and reading comprehension skills. It defines efficiency as real classroom economy in development, administration, and scoring without compromising validity, reliability, or fairness.

2.1 Theoretical and Conceptual Foundations of Language Assessment

This study is based on the "*test usefulness*" perspective (Hughes (a), 2003). This perspective looks at the quality of an assessment as a combination of validity, reliability, authenticity, interactiveness, impact (washback), and practicality (Brown & Abeywickrama, 2010). According to this perspective, different authors such as Baxter (1997), Brown (2004), and Hughes (a) (2003) have supported the principles of language assessment for assessing what is intended to be assessed appropriately. The two main principles of assessment are *reliability* and *validity*. That *reliability* means to what extent the assessment provides a consistent result, also *validity* means to what extent the assessment process assesses what is intended to be assessed (Brown, 2004). Meanwhile, giving feedback, which is '*washback*' is the effect of assessment on the teaching and learning process. This principle of assessment will show the impact

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

on both teachers and students, on the educational system, and on society in general (Hughes (a), 2003).

Another two important principles are *practicality* and *efficiency* in which both focus on the idea that the assessment should not be a time-consuming and an expensive process (Baxter, 1997; Hughes (a), 2003; Oller Jr, 1979). *Practicality* addresses feasibility within real constraints of time, budget, logistics, and teacher workload (Baxter, 19973; Brown & Abeywickrama, 2010; Oller, 1979); *efficiency* can be viewed as the practical ratio of quality to resources across development, administration, and scoring (Bekleyen, 2010; Carroll, 1973; Weir, 1991). Bekleyen (2010) has explained that an important term which is related to practicality and cost is efficiency. Weir (1991) have identified it as a quality of a test. Also, Carroll (1973) has stated that one of the problems that can occur in language assessment construction and administration is the problem of efficiency, which is related to limited time, money, and availability of resources for the language testing process. Furthermore, *authenticity* is another principle that relates to validity because it emphasizes that to which extent an interpretation or an assessment result can correspond with the real language use (Bachman & Palmer, 1996; Brown and Abeywickrama, 2010).

Bekleyen (2010) has stated that practicality and efficiency can be operationalized with concrete, specification-linked indicators so that "saving time" never comes at the expense of score meaning. Practicality and efficiency in assessment go beyond time savings, cost-effectiveness, resource management, and alignment with instructional use to preserve score meaning and validity. During the test development process, indicators such as hours from blueprint to pilot-ready form, the number of review cycles, and the proportion of items accepted at first pass allow developers to monitor efficiency without compromising construct coverage (Crocker & Algina, 2006; Downing & Haladyna, 2011). While minimizing the administration time for efficiency, the test length and task complexity should be balanced to achieve the target reliability and validity (Crocker & Algina, 2006; Bachman & Palmer, 1996). During the pilot testing process, the developers can support efficiency by revealing unrelated and incorrect items, non-functioning distractors, or tasks that allow for revisions before operational use (Downing & Haladyna, 2011; Fulcher & Davidson, 2007). Finally, efficiency in scoring and feedback is important: prompt scoring, reporting results, and making decisions about the results (Brown & Abeywickrama, 2019; Bachman & Palmer, 1996). This process means that test should be scored **quickly**, after that the results should be **clearly reported**, and the scores should be **used to**

make decisions. When scoring is done promptly, teachers can give feedback on time and use the results to improve teaching and learning.

2.2 Reading as a Process and Product in EFL context

Reading can be conceptualized both as a process—an interaction in which readers decode, parse, and integrate textual information with background knowledge—and as a product, observable in comprehension outcomes such as accurate inference, identification of main ideas, and cohesion tracking (Weir, 2005). In addition, Wahyudin et al. (2024) have explained that reading skill is focused on in the English foreign language contexts. Further, this process goes through encoding written information, interpreting, and comprehension of that information, which is the product of the reading process. The product model of reading emphasizes what a student understands from a text instead of how the student understands the text (Macmillan, 2016; Weir, 1998). Brown (2004) has stated that reading skill is an essential skill that students can develop naturally. Also, he has explained that at the basic level, the textbooks are designed in a way that focus on the learners' reading ability. Skilled readers rapidly recognize frequent words and common syntactic patterns, freeing working memory for integration and inference, whereas less skilled readers devote more resources to lower-level decoding and have less capacity for building a situation model (Brown, 2004; Grabe, 2009).

2.2.1 Testing Reading Comprehension: Techniques for Measuring Process and Product

Assessing reading comprehension in EFL contexts requires a thoughtful combination of techniques that capture both the process and product of reading. Different techniques and methods can be used to measure the ability of the reading process and product. Alderson (2000) has stated that there is no only one method or technique for measuring reading skills that will fulfill the intended purpose of the teacher for testing, but there are certain methods and techniques that can be used, such as cloze questions and multiple-choice items. Also, Heaton (1990) has argued that, instead of spending most of the time just using one item type, such as multiple-choice, which takes much time and effort, teachers can use different item types, such as open-ended questions, those questions that require the test-taker to use their own words and ideas. In classroom contexts, the teacher's choice of technique depends on the instructional goals, learners' proficiency levels, and the specific subskills to be measured (Stiggins, 1992; Stiggings et al., 2004; Swaie & Algazo, 2023). Hence, effective test design balances practicality with validity by using varied item types such

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

as cloze tasks, multiple-choice questions, and open-ended responses to provide a more comprehensive picture of the learner's reading ability.

Alduais (2013) has mentioned some test formats that can be used for testing reading, and the common format that can be used for assessing the students' readability is cloze test format, and the other formats are matching pictures with sentences, true/false, finding correct sentences, completion items (word, phrase, etc.), short answers for WH-questions, and rearranging. In addition, Heaton (1990) and Hughes (b) (2003) have emphasized the importance of selecting techniques appropriate to the text type and testing objective. They have suggested a combination of item formats—multiple-choice, true/false, matching, reordering, and open-ended questions—to enhance content validity and maintain learner engagement.

2.3. Chatbots and Conversational Agents in Language Education

Integration of AI tools such as Chatbots in the language teaching and learning process modifies how teachers and learners interact with the process, in which it is a computer application that imitates human conversations (Danesi, 2024; Gutiérrez, 2023; Kohnke et al., 2023). Chatbots create an immersive environment for the learners (Chen et al., 2024). AI supports the process by allowing the learners to interact with the learning materials, providing more personalized learning, and improving language speaking and listening skills (Ayala-Pazmiño and Alvarado-Lucas, 2023). Chatbots such as GPT, Gemini, and Cloude can process natural language through texts and images (Galczi & Luckin, 2024). Furthermore, Singha et al. (2024) have stated that AI Chatbots are conversational agents that help learners to interact in a dialogue by providing immediate feedback and correcting mistakes, such as Duolingo and Microsoft Xiaoice, which provide real-world language practices for the users to feel they are interacting with a real human and having a real conversation. Chatbots use the three AI technologies for providing better outcomes, which are Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) (Gutie`rrez, 2023). The use of Chatbots in education dates back to the use of the ELIZA program, which is a text-based conversation program that mimics human conversation (Koraishi, 2023). Gutiérrez (2023) has stated that chatbots can be used for diverse purposes through having conversations with an AI agent, such as for practicing language, because they imitate human language through providing real conversations. Also, it can be used for providing instance feedback to the users or students and finding their mistakes after completing a task to use the correct form of language (Yatri et al., 2023). Teachers and instructors use chatbots such as ChatGPT for preparing

tasks (Yatri et al., 2023). Also, ChatGPT is used for text summarization, creating content, question answering, and other tasks (Johnsson, 2023). Additionally, chatbots use text, image, and voice to increase students' engagement with the tasks and learning outcomes. The use of chatbots helps learners to improve their writing errors through providing opportunities for writing practices, and writing practices, and conversational agents help students communicate based on their interests (Bailey, 2019).

AI-powered chatbots have a useful role in language-test design by serving as **test-item generators, assessment assistants, and personalized practice**. According to Alsagoafi and Alomran (2024) teachers use ChatGPT in language test design to draft tests, producing quicker, more diverse, and more detailed types of questions to minimize workload and save time. More broadly, Wiboolyasarin et al. (2025) have explained that AI chatbots support language skill development, and this benefit helps teachers to generate test **questions** for language skills, such as generating questions for reading comprehension passages, grammar questions, and writing prompts for writing or speaking tasks. Moreover, chatbots support **personalized assessment design** by focusing on learners because of their ability to interact in natural language and respond adaptively. Chatbots help students to be assessed at their own pace and level (Li et al., 2025). In addition, Koç and Savaş (2025) have highlighted that Chatbots produce a positive environment with low-anxiety learners and make them feel confident, which is beneficial for low-level students or for learners who may feel a lack of confidence in traditional classroom assessment. Finally, using chatbots in test design **increases efficiency and scalability,** in which automated test design and quick item generation reduce the teachers' manual time and effort for test design, especially when it is useful for large classes or limited teacher resources (Hadzhikoleva et al., 2024).

2.3.1 ChatGPT's benefits and roles in Language teaching and learning

Chat GPT is one of these chatbots that has reached one million users in only five days after its launch (Thao, 2023). Chat GPT is a language model that stands for Chat Generative Pre-trained Transformer, which was released in November 2022 and developed by OpenAI Company that enables users to engage in a natural conversation with the system (Dergaa et al., 2023). According to Mbwambo and Kaaya (2024), this language model is easy to use by users. The evolution of ChatGPT opens ways for developing other generative AI chatbots, such as Gemini, which is a

Google chat service that is known as Bard, Bing Chat of Microsoft, and LLaMa of Meta (Hsiao et al, 2023).

Thao (2023) has explained that the chatbot helps both teachers and students in their learning and teaching foreign languages because of its ability to create an interactive, engaging, and productive context for developing language skills, also it offers a personalized and adaptable learning experience (Ma et al., 2024). Jeon and Lee (2023), have identified four main roles of ChatGPT in education, which are assistant, interlocutor, material supplier, and assessor. According to Thao (2023), it can be used for improving and developing the reading and writing skills through providing written materials that reflect the real-world contexts; however, it has a neutral impact on developing listening and speaking skills, but still, if students can interact with the voice interactions regularly, they can improve their listening ability and pronunciation. Meanwhile, it can be used for developing grammar, vocabulary knowledge, and translation between languages (Ma et al., 2024; Thao, 2023).

Harunasari (2023) have mentioned some uses of ChatGPT, which can be used for researching, translating, paraphrasing, and summarizing. Meanwhile, Kasneci et al. (2023) through their study on ChatGPT, they have reached some potential applications of using it in education, in which it can be used for facilitating reading and writing skills, and language learning generally. ChatGPT helps students in their research writing process and assists them with finding and addressing a variety of questions around the research topic (Kasneci et al., 2023). Also, it develops students' critical thinking and problem-solving skills, creates summarization and finding answers for complex tasks, and increases engagement through facilitating group debates and discussions (Suriano et al., 2025). It helps the students with disabilities through the use of both speech-to-text and text-to-speech features (Holmes et al., 2019). According to Kasneci et al. (2023) it helps teachers and students in the assessment process through designing quizzes and exams, grading students' responses, and providing self-feedback, it provides the basic information around topics, works as a personal tutor and encourages personalized learning, provides guidance for designing teaching methods, lesson planning, and assessment process for teachers, and suggests learning materials for students based on their needs.

2.3.2 AI Assistance and ChatGPT in Reading Development and Reading Test Design Process

Designing a high-quality test that has the characteristics of a useful language test, especially in terms of validity, will be time-consuming, and ensuring its quality is a complex process that requires knowledgeable and experienced teachers (Säuberli and Clematide, 2024). Various AI-driven tools support the enhancement of reading comprehension. The AI tools support reading through some functions, such as translation, speech-to-text, and text-to-speech, help students to decode, understand, and remember the academic texts through using NLP and ML algorithms (Zafar, 2025). SARA is an AI reading assistant for reading comprehension that integrates both Large Language Models (LLMs) and Eye Tracking systems for enhancing comprehension and improving reading skills, which helps learners to overcome the difficulties in unknown words and complex sentences (Thaqi et al., 2024). Another platform, which is identified by Zyska et al. (2023), is CARE stands for (Collaborative AI-Assisted Reading Environment), which enhances reading comprehension through using NLP algorithms, classification of texts, and question answering. This platform allows the readers to read and discuss the texts collaboratively with the help of AI, which means the users can comment and highlight the text (called inline commentaries), and it can automatically provide more information on their comments and highlighted sections.

Alsagoafi and Alomran (2024) have highlighted that teachers use ChatGPT in language test design to draft tests, producing quicker, more diverse, and more detailed types of questions to minimize workload and save time. More broadly, Wiboolyasarin et al. (2025) have explained that AI chatbots support language skill development and that these benefits help teachers to generate test questions for language skills, such as generating questions for reading comprehension passages, grammar questions, and writing prompts for writing or speaking tasks. In addition, Koç and Savaş (2025) have highlighted that Chatbots produce a positive environment with low-anxiety learners and make them feel confident, which is beneficial for low-level students or for learners who may feel a lack of confidence in traditional classroom assessment. The most important use of chatbots in test design is increasing efficiency and scalability, in which automated test design and quick item generation reduce the teachers' manual time and effort for test design, especially when it is useful for large classes or limited teacher resources (Hadzhikoleva et al., 2024). According to Chun and Barley (2024), Säuberli and Clematide (2024), and Shin and Lee (2023), ChatGPT can

be used for creating multiple-choice items for reading comprehension skills that can be effective as those created by expert teachers. According to O (2024), the chatbot can be used for designing true/false question items for reading. Chun and Barley (2024), Säuberli and Clematide (2024), Shin and Lee (2023), and O (2024) support that the chatbot can be used for generating reading question items based on well-organized prompts. Also, Ma et al. (2025) have explained that ChatGPT can be used for creating inference-making reading comprehension assessments, which is an essential skill in reading comprehension tests that depends on the students' prior knowledge.

2.4. Related Studies

The growing interest in large language models LLMs in language education and the assessment process leads many studies to work on the issue and produce some interesting results. In recent years, there has been growing interest in investigating the efficiency and capability of AI, particularly ChatGPT, in designing reading comprehension tests compared to human experts. There are studies across different contexts that have examined the capability of AI tools in generating reading test items, especially multiple-choice and reading passages, and compared them directly with those designed by humans. Shin and Lee (2023) have investigated whether ChatGPT can design reading comprehension passages and multiple-choice test items compared to those that are designed by human experts or not, which ChatGPT was used to design 5 reading passages with multiple choice items that are equal to the test designed by human. The data were collected through using Likert-scale questions and open-ended responses from **50 participants** who were **pre- and in-service English teachers** to rate the question items based on these criteria; naturalness of the passage, expression quality, attractiveness of multiple-choice items, and overall quality of the items. They have found that the passages that were created by ChatGPT were as natural as those created by humans, helped teachers design tests in a short period of time, and reduced their workloads, but still, human intervention was necessary for creating answer options.

Sihite et al. (2023) have examined the capacity of ChatGPT in designing reading comprehension questions for academic texts, with the focus of the study being on questions' alignment with higher-order cognitive skills. The test questions, which are 30 items, were generated by ChatGPT based on three selected TOEFL ITP reading comprehension passages, in which each passage contains 10 items. The study sampled **25 university students that participated in the test.** The results have indicated that ChatGPT is capable of producing questions that cover a

range of cognitive levels, but only 10 items met the validity criterion, and the reliability of items was moderate (.671), which suggests a reasonable level of consistency.

Additionally, Chun and Barley (2024) have explored the potential use of the free version of ChatGPT that was used instead of the paid version because of teachers' access limitations. The chatbot was used to improve the efficiency of the test development process. The researchers compared two tests for designing 80 multiple-choice items that 40 were created by teachers and 40 by the chatbot, all based on **20 authentic Korean passages**. The item quality was evaluated by three expert raters through using **five-point Likert scale** with a rating **rubric for evaluating item quality based on eight criteria**, and providing written comments. The results have indicated that the items written by ChatGPT are clear and grammatically correct, but the chatbot's ability to design multiple-choice distractors is limited, and there is a need for human judgment in choosing appropriate distractors.

A comparative study by O (2024) has compared both human-made and AI-made tests to examine the efficiency of ChatGPT in designing test items alongside human-made items. Each of the tests consists 20 items that one designed by a human, another one designed by ChatGPT. The test contents are designed to assess students' achievements in the course, which the test items consists of only 10 multiple-choice and 10 true/false items. The sample is 20 university students enrolled in a TESOL theory course. In addition, the results have indicated that ChatGPT could design test items comparable to human-created ones in linguistic quality and structure, it is confirming AI's efficiency and capability in test generation and assisting teachers in test creation, saving their time and reducing workload.

Kanık (2024) also have used ChatGPT in designing tests to make a comparison between the items generated by teachers compared to those by the AI tool to find whether the multiple-choice items designed by ChatGPT can be effective and in quality and usefulness compared to human-made ones. The tests were given to a group of students who are 36 university students, the tests have consisted of 40 items 20 designed by teachers, and 20 by ChatGPT. The author has found that the AI tool could design items that have acceptable level of item discrimination, and the tool is effective in generating items, providing contextual-richness, and consistency in meaning-focused skills.

Meanwhile, Shin et al. (2025) have examined the efficacy of AI in designing reading tests by comparing them with human-made tests. The study used expert reviewers, raters, and **English language specialists**

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

(without explicitly specify the number of participants involved in the study) to rate the **linguistic quality, structural features, and test design characteristics of the test items**. The results have indicated that ChatGPT could design test items comparable to human-created ones in linguistic quality and structure, it is confirming AI's efficiency and capability in large-scale test generation.

Akpan (2025) has examined the **capabilities of large language models (LLMs)**, specifically **ChatGPT, Claude, and Gemini against human literacy and educational benchmarks. The study was not experimental study** in the classroom, the tools are compared systematically in test design process against humans. The findings have indicated that the tools are performed well compared to humans in designing reading test items, also outperformed better than humans in terms of efficiency, consistency, and linguistic clarity. Another key finding was that **AI models outperformed human benchmark performance** in **advanced reading comprehension**.

Furthermore, there are researchers that have explored the teachers' perceptions towards the use of ChatGPT in language exam creation. Jeon and Lee (2023) have explored the potential role of ChatGPT in language education and how the use of the chatbot can change the teachers' roles. The participants were **11 language teachers** who were asked to use **ChatGPT in their teaching activities** for about **two weeks**. After the two-week period, each teacher participated in an individual semi-structured interviews to explore teachers' perceptions, experiences, challenges, and perceived benefits of using ChatGPT in their teaching. The study's results have showed that it enhanced the teachers' roles, but did not replace them, and that it can work as an assistant, interlocutor, content provider, and evaluator in language classrooms.

Similarly, Alkhateeb et al. (2025) have conducted a study to explore **61 EFL teachers' perceptions** from **two universities** who participated in an **online survey** questionnaire about the use of AI tools such as ChatGPT, Cloude, Gemini, and Exam Software in preparing exams with the focus on efficiency, benefits, and challenges of AI. The study have indicated that teachers use ChatGPT mostly for generating objective test items such as multiple-choice, true/false, and fill-in items.

3. Methodology

This chapter outlines the methodological framework employed for finding the objectives of the research. The methodology includes an explanation of the research design, participants, instruments, data collection tool and procedure, scoring scheme, test content validity, and ethical considerations. The inclusion of student performance data

provides a comprehensive understanding of the role and potential of AI tool ChatGPT in educational assessment.

3.1 Research Design

This study employed a **quasi-experimental research design** to compare the results of two English reading comprehension tests that one designed with ChatGPT and another one designed by humans to evaluate differences in students' performance, additionally, to show if ChatGPT generated tests are efficient enough for EFL teacher to use in designing tests.

3.2 Participants

The sample of the study were a group of students at the University of Halabja, English language department, second stage for the academic year 2024-2025. The participants' age was between 19-20, and their native language is Kurdish. The number of students was 51, of which 38 female and 13 male students who have all participated in both tests AI generated and teacher made reading comprehension tests.

3.3 Instruments

The data were collected through two designed reading comprehension tests, one is teacher-made test and the other one designed by ChatGPT (see Appendix 1). **After** the tests were ready, they were given to a group of university EFL students. Each student took both the AI-generated and teacher-made tests. All students were given the same instructions and time limits (60 minutes for each test), and the tests were taken in the same kind of environment, which was administered in a classroom to avoid any outside influence on performance. Importantly, students did not know which test was AI-generated and which was teacher-made to avoid bias.

The teacher-made test consists of 12 items, which were designed by the researchers in collaboration with the module instructor. The test focused on reading comprehension skills such as scanning, skimming, finding the main idea, grammatical knowledge, vocabulary knowledge, guessing, and activating students' critical thinking. Out of 12 items, 2 of them tested grammatical knowledge, and 4 items tested vocabulary knowledge. The last 6 items tested the students' comprehension skills. The AI-assisted test also consisted of 13 items, which were designed by ChatGPT. Similar to the teacher-made test, the AI-generated test focused on the same reading skills, which were vocabulary, grammar, and comprehension skills. The AI comprehension items have focused on similar comprehension skills as teacher-made items, with three more skills, which are activating background knowledge and inference, recognizing purpose and hidden information, and understanding text structure.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

3.4 Data Collection Procedure

Quantitative data were collected from two sources: students' performance on the two English reading comprehension tests, which were an AI-designed test and a teacher-made test. Both tests were administered to the same group of undergraduate EFL students inside their classroom. Each test lasted 60 minutes and was supervised by the course instructor and researchers to ensure consistency. The tests were administered for students' midterm exam to have reliable data and make the students answer accurately.

3.5 Scoring scheme

The items were scored manually by the researcher and the module instructor. The procedures for scoring both subjective and objective items were different that the objective ones have one correct answer, and the scores will be zero or a full mark. While the subjective items which included the comprehension questions, the focus was not on the grammatical errors of students; the focus was on students' understanding and reading ability of the texts and giving the ideas from the text for each test item. The criteria that were used for scoring both tests were based on the students' grammatical knowledge and vocabulary (for the items that are directly focused on these two skills), word choices, finding the main idea, activating students' critical skills, scanning, skimming, and so on. The collected data were coded, organized, and prepared in SPSS for subsequent statistical analysis.

3.6 Test content validity

Test validity was used to ensure the quality of both tests. Both tests were given to a group of expert members which consists of university teachers to ensure face and content validity. A group of 13 expert EFL instructors participated as jury members to evaluate the content validity of the two reading comprehension tests. All experts held MA or PhD degrees in Applied Linguistics or English Language Teaching, with professional experience ranging from 7 to 20 years in teaching EFL at the university level. Their academic background in language assessment and prior involvement in test construction qualified them to judge item clarity, relevance, linguistic accuracy, and alignment with reading comprehension skills.

3.7 Ethical Considerations

This study follows standard ethical guidelines and principles, which is outlined in both the Belmont Report (1979), the British Educational Research Association (BERA) Guidelines (2018), and Research Design: Qualitative, Quantitative, and Mixed Methods Approaches by Creswell & Creswell (2018) in which these sources emphasized being justice towards

the participants, respecting them, and beneficence. The participants of the study were university-level English language students. Students' identities were kept anonymous through encoding their names to ensure anonymity and confidentiality. There were no physical or psychological harms to participants. The results of the data were reported in a way that prevents identification of the participants to ensure trustworthiness and protect the participants' rights. Before collecting data from the participants, the researcher asked the head of the English department and the instructor of the lesson for permission to do the work, and it was accepted by both the head of the department and the instructor with the agreement to follow legal and institutional rules.

4. Data Analysis and Results

The objective of this study was to investigate whether there is a significant difference between the efficiency of AI-designed language reading comprehension tests and teacher-made tests across three key language areas and skills: grammar, vocabulary, and reading. To address this research question, a repeated measures analysis of variance (RM ANOVA) was conducted to compare the performance of students on both test types (ChatGPT-designed and teacher-made) across the three skill areas. Both descriptive statistics and RM ANOVA were used to analyze the data. The descriptive statistical tools are Mean (M), Standard deviations (Std) for each reading skills. M is the value that is calculated in averaging something (in a mathematical term it is the sum of scores divided by the number of scores (Fraenkel et al., 2012; Navarro, 2019). Also, SD is a number that shows how spread out or how close together the data values are around the mean (average) (Navarro, 2019).

Additionally, RM ANOVA referred to as within-subjects ANOVA was a statistical test to find whether there is a significance differences between related means, and to find the differences of a dependent variable in different situations, conditions, or time. In this study it provided the outcome which is multivariate analysis of variance that is used for analyzing the multiple variables simultaneously and finding their effect on other variables, and allows to look at the link between the variables (Hamad, 2025). This multivariate test is important for assessing the main effect on each independent variables and the interaction effects between them, and effect of one variable on one or more dependent variables (Hamad, 2025; Mengual-Macenlle et al., 2015). Here, it is used to examine whether there are significant differences in students' performance (dependent variable) and finding the effects of the test types, the skills, and their interaction (independent variables) on students' performance.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

4.1 Descriptive Statistics

The descriptive statistics of the key areas of language, including grammar, vocabulary, and reading skills, are presented in Table 1 Descriptive Statistics. The results indicate that students performed differently across the two types of tests, as well as across the three skills. Specifically, the mean score for the AI-designed test was higher than the teacher-made test in the areas of vocabulary and reading, whereas the teacher-made test had slightly higher mean scores in grammar. For the AI-designed test, the mean scores were .59 for grammar, 2.86 for vocabulary, and 9.84 for reading, with corresponding standard deviations of .92, 1.50, and 3.13, respectively. In contrast, the teacher-made test yielded mean scores of 1.02 for grammar, 1.08 for vocabulary, and 9.49 for reading, with standard deviations of .79, 1.06, and 3.37, respectively. These mean values suggest that the AI-designed test generally yielded higher scores in vocabulary and reading, indicating its greater efficiency in assessing these skills. However, the teacher-made test scored slightly higher in grammar, suggesting that it might be more effective for evaluating grammar comprehension.

Table 1

Descriptive Statistics

AI-designed Test						**Teacher-made test**					
Grammar		Vocabulary		Reading		Grammar		Vocabulary		Reading	
Mean	Std. Dev	Mean	Std. Dev	Mean	Std Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
.59	.92	2.86	1.50	9.84	3.13	1.02	.79	1.08	1.06	9.49	3.37
Total											51

4.2 Multivariate Tests: Main and Interaction Effects

The multivariate tests (See Table 2) revealed a significant main effect for test type, with an F-statistic of 13.478 and a p-value of .001, which is statistically significant. This result confirms that there is a significant difference between the AI-designed and teacher-made tests overall. The effect size for this difference, measured by partial eta squared .212, suggests that the test type accounts for about 21.2% of the variance in

student performance. This indicates a moderate effect of test type on the results, with the AI-designed test generally being more efficient in assessing students' language skills.

Furthermore, the analysis showed a highly significant main effect on language areas and skill, with an F-statistic of 272.611 and a p-value of .000. This result confirms that students' performance differs significantly across the three skill areas—grammar, vocabulary, and reading. The effect size for skills, measured by partial eta squared .918, is very large, indicating that the skill type is the dominant factor influencing the variance in test scores. This suggests that different skills require different levels of attention and may be assessed with varying degrees of effectiveness by the two test types.

Table 2

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Test_Type	Pillai's Trace	.212	13.478	1.000	50.000	.001	.212
Skills	Pillai's Trace	.918	272.611	2.000	49.000	.000	.918
Test_type * skills	Pillai's Trace	.604	37.404	2.000	49.000	.000	.604

Finally, the analysis revealed a significant interaction effect between test type and skill type, with an F-statistic of 37.404 and a p-value of .000. The significant interaction effect between test type and skill type further emphasizes that the difference between the AI-designed and teacher-made tests is not uniform across all skills. For grammar, the teacher-made test produced higher scores, while for vocabulary and reading, the AI-designed test yielded better results. This indicates that the AI-designed test is more efficient in evaluating vocabulary and reading, while the teacher-made test may be better suited for assessing grammar.

Overall, the results provide clear evidence that the AI-designed test is generally more efficient than the teacher-made test in evaluating students' language reading comprehension, particularly in the areas of vocabulary and reading. However, the teacher-made test outperforms the AI-designed test in grammar, suggesting that the two test types have strengths in different areas of language assessment. The interaction effect indicates that the difference between the two test types is not consistent across all

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

skills, emphasizing that the effectiveness of a test may depend on the specific language area and skill being measured.

5. Discussion of Results

This research paper aims to examine the efficiency of using ChatGPT in language education for designing English as a Foreign Language (EFL) reading comprehension tests compared to the traditional method of test design, which is a teacher-made test. The researcher compared the students' performance on ChatGPT-generated test items with the traditionally teacher-made reading comprehension tests. Especially, to find whether there are any significant differences between students' performance on both tests across grammar, vocabulary, and reading comprehension skills. The results revealed statistically significant differences in students' performance across both tests, and across all the skills which students performed better in the ChatGPT-designed test in both vocabulary and reading, but performed well in grammar in the teacher-made test. These results provide valuable insights into the effectiveness of ChatGPT in designing reading tests across vocabulary and reading comprehension skills, and the skill-specific impact of AI-generated tests on EFL learners' performance. The results of this study supported by both Kanık (2024) study that the AI tool ChatGPT is more efficient and effective in designing reading tests, because it can provide benefits in terms of item generation, contextual richness, and consistency in meaning-focused skills in language assessment.

The results are in contrast with the previous studies' results, which suggest that ChatGPT is effective, comparable to human-made, not superior to it, especially in designing vocabulary and reading skills (O, 2024; Shin & Lee, 2023; Shin et al., 2025; Sihite et al., 2023). Instead, the findings are in line with Akpan (2025), who argued that LLMs such as ChatGPT, Gemini, and Cloude significantly outperform human benchmarks in tasks such as undergraduate knowledge and advanced reading comprehension. The main factor that may lead to these differences in the results between the current study and other previous studies is that this study focuses on whether the AI tool ChatGPT could be effective and efficient in designing reading skills. Also, it focused on assessing ChatGPT's capability in functioning across reading skills through comparing students' performance in the tests across the reading skills. However, the previous studies did not assess how ChatGPT is capable across the reading skills that they often treated reading as one single skill or a unified skill (Shin & Lee, 2023; Shin et al., 2025). Much of the existing studies have examined the quality of ChatGPT-generated test items or the general capabilities of AI systems, rather than comparing

students' performance on both tests (Chun & Barley; Shin & Lee, 2023; Shin et al., 2025). Instead, the study by Sihite et al. (2023) has focused on students' performance on reading comprehension items that also focused on comprehension skills. At the same time, Sihite et al. (2023) has not focused on vocabulary and grammar to be embedded in the tests, and did not compare students' performance with a traditionally designed test. Furthermore, another factor leads to this difference because of the previous studies' focus on teachers' perception in comparing the AI-ChatGPT-generated test to human-made tests in terms of item qualities without focusing on students' authentic performance on both tests (Alkhateeb et al., 2025; Jeon & Lee, 2023).

One of the most salient result from the descriptive statistical analysis indicated that there is a significant difference between the tests, which means students performed differently across both tests and across the skills. The results of the current study indicated that students performed well across both vocabulary and comprehension items in the ChatGPT-designed test, compared to the teacher-made test. This indicates that ChatGPT is capable of designing vocabulary and reading comprehension items better than human-made. In contrast, the study by Shin et al. (2025) have demonstrated that the AI-designed reading items faced challenges in vocabulary control and distractors' effectiveness. The mean scores and the standard deviation of the ChatGPT-generated vocabulary items are (M = 2.86, SD = 1.50) and the teacher-made test (M = 1.08, SD =1.06). Additionally, the mean and standard deviations of ChatGPT reading comprehension items is (M=9.84, SD= 3.13) and the teacher-made test is (M= 9.49, SD= 3.37). The mean and standard deviations of the ChatGPT-designed vocabulary and reading items are higher than the teacher-made test, which indicates that students performed better and showed more variation in the ChatGPT-designed vocabulary and reading items. These results show that students' performance varied as some did better than others and did not have similar marks in both. Additionally, the results support that the scores in the ChatGPT test are more spread out, which indicates that the test is better at distinguishing the students' level of performance. In other words, the ChatGPT test was able to show real differences in how well students understood the materials. This kind of spread is useful in educational testing because it has helped to identify both stronger and weaker students and suggests that the test was neither too easy nor too difficult for students.

One of the factors that may lead to these significant differences in vocabulary items across both tests, in which a stronger performance was observed in ChatGPT's vocabulary items, can be attributed to the AI

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

systems that are trained on a large amount of lexical and vast textual data. This can enable AI technologies such as ChatGPT to generate a large amount of vocabulary data within a short period of time (Cong, 2024; Kalyan, 2023; Panagiotidis, 2024). This ability of ChatGPT may lead to the creation of items that can help students access word meanings more efficiently during the testing process. In contrast, teachers' limited time and effort may limit their abilities in designing high-quality vocabulary items; also, the teacher may depend on their personal experience, curriculum materials, and textbooks, which can limit the range of vocabulary items to be included in test items and reduce contextual richness that could affect students' performance. Similarly, the higher performance observed in ChatGPT-designed reading comprehension items may have happened because of several interrelated factors that are associated with the nature of AI-generated texts and questions. Because of AI systems' capability in designing coherent and logically structured reading texts and passages that are organized clearly, this helps learners not to be confused by comprehension passages, which is supported by Shin and Lee (2023). Furthermore, the ChatGPT-designed reading comprehension items may have reflected authentic language use better than those designed by teachers because of making links between natural discourse patterns and varied textual features, that ChatGPT has this ability to use authentic language (Thotad et al., 2022). As a result, exposure to authentic input can enhance students' engagement and facilitate a deeper understanding of text meaning.

However, in grammar items, students performed better in the teacher-made than in the ChatGPT-designed test based on the descriptive statistics. The mean and standard deviations of ChatGPT test is (M=.59, SD= .92), and the teacher-made test is (M= 1.02, SD= .79). The results of the mean and standard deviations of teacher-made grammar scores indicate that students performed better on the grammar items developed by teachers. The findings show that students experienced greater difficulty as well as substantial variability in their performance in the ChatGPT-designed grammar item. This suggests that the ChatGPT-designed grammar items may not have been aligned with students' instructional background, proficiency level, or the specific grammatical structures that are emphasized during classroom instruction. While some students were able to answer the grammar item correctly, others struggled with the grammar, which led to inconsistency in their performance and wider score distribution. In contrast, the higher scores in the teacher-made grammar test indicate that students performed better and stronger and provided higher consistency in students' performance. This result may

show that teacher-designed grammar items were more closely aligned with instructional content, classroom practice, and learners' actual grammatical proficiency level. This supports that teachers are able to design grammatical items that can reflect what has been taught accurately and capture students' level of performance in grammar. There may be key factors for having these results which can be linked to several pedagogical and contextual factors. Teachers mostly design the grammar items that are aligned with the instructional materials that are directly from classroom instruction, textbooks, and lesson objectives, which reduces their confusion (Khan et al., 2025; Martone &Sireci, 2009). So that, students' familiarity and teachers' repetition of the item formats, or common phrasing patterns, all of this may lead to get higher scores in teacher-made tests. Furthermore, teachers' contextual and cultural awareness, which enables them to avoid unnecessary linguistic complexity and those items that do not match learners' proficiency levels, may have increased students' grammar scores. Even if ChatGPT-generated grammar items are linguistically accurate, they may not always reflect students' instructional background, which leads to lower overall performance. As a result, these factors may provide a more accessible and supportive testing environment, which teacher-made grammar items can lead to higher performance compared to ChatGPT-designed grammar items.

The results of multivariate analysis using Pillai's Trace in Table 2 revealed that the effect of test type, the skills (vocabulary, grammar, reading comprehension), and the interaction between test type and reading skills on students' performance is significant. First, statistically significant main effect of test type (F-statistic of 13.478 and a p-value of .001) shows that students' overall performance between the ChatGPT-designed and teacher-made tests differed significantly. It means the test types, GPT-designed and teacher-made tests, affected students' performance and the effect is significant. **Additionally**, the **main effect of skills on students' performance (**F-statistic of 272.611 and a p-value of .000) **indicates that students' performance differed across the skills (vocabulary, grammar, and reading) without depending on the test types. Furthermore,** the significant interaction effect between test type and skills (F-statistic of 37.404 and a p-value of .000) emphasizes that the difference between the ChatGPT-designed and teacher-made tests is not uniform across all skills. **This means students performed in each skill differently in the same test that there may be a student who got a high score in vocabulary, but got a low score in grammar or comprehension skills regardless to the type of the test who designed**

it. Furthermore, the significant interaction effect between test type and skills emphasizes that the difference between the ChatGPT-designed and teacher-made tests is not uniform across all skills. This means that the ChatGPT-designed test produced higher scores in vocabulary and reading, but the teacher-made test was more effective in designing grammar. These findings demonstrate that the effectiveness of assessment does not only depend on the mode of test construction but also on the specific language skill being measured.

These results provide important insight into the influence of test type, language skills, and their interaction on students' performance. The partial eta squared for test type (.212) demonstrates a moderate effect, which means both tests, whether ChatGPT-designed or teacher-made, have a meaningful influence on students' overall performance across the three reading comprehension skills. In contrast, the effect of language skills was large (.918), which indicates that students' performance differs substantially across grammar, vocabulary, and reading, and confirms that skill type is the dominant factor shaping outcomes. Furthermore, the interaction between test type and skills also showed a large effect (.604) that highlights the relative effectiveness of ChatGPT-designed and teacher-made tests is not uniform across skills; instead, it depends strongly on the specific language area being assessed, which means **ChatGPT-designed test is not always better than the teacher-made test**, and the **teacher-made test is not always better than the ChatGPT-designed test**, but their effectiveness **depends on which language skill is being measured**. It means both tests, whether ChatGPT-designed or teacher-made, have a meaningful influence on students' overall performance across the three reading comprehension skills. In contrast, the effect of language skills was large, which indicates that students' performance differs substantially across grammar, vocabulary, and reading, and confirms that skill type is the dominant factor shaping outcomes. This raises important questions about the role of AI tools such as ChatGPT in language assessment, about whether AI tools should be used across skills uniformly or be used only for the skills where it can perform best. The findings suggest that AI tools is most effective to be used to support the test items that focus on meaning-oriented skills, but grammar test items require human involvement substantially.

Overall, the results indicate that the effectiveness and efficiency of a test depends on both the type of test and the specific language skill being assessed, highlighting the importance of using assessment approaches that are sensitive to skill-specific needs and that combine the strengths of

ChatGPT-generated tools with teacher expertise and these findings highlight the importance of teacher expertise involvement in grammar test design and teacher judgements about the item and ChatGPT-designed grammar items should be used cautiously by teachers and reviewed to ensure instructional alignment, appropriate difficulty levels, and reliable measurement of learners' grammatical competence. This supports a complementary approach, in which AI tools may enhance assessment efficiency for skills such as vocabulary and reading, while teacher expertise remains crucial for form-focused evaluation of grammar. The findings are in line with Chun and Barley (2024) and Shin and Lee (2023), who also supported the need for teachers' involvement in the testing design process for giving a better outcome. These results indicate that ChatGPT or other AI tools should not be viewed as a replacement for teachers in designing language tests, but should be considered as a complementary tool that can assist teachers in enhancing test design efficiency and effectiveness when used appropriately.

6. Conclusion

Depending on the results analyzed, the following conclusions are arrived at:

1. This study supports that the AI tool ChatGPT is more efficient and effective in designing reading tests, because it can provide benefits in terms of item generation, contextual richness, and consistency in meaning-focused skills in language assessment.
2. AI tool ChatGPT is more efficient in designing meaning-oriented skills instead of form-focused skills.
3. The ChatGPT-designed test was effective in assessing vocabulary and reading comprehension because of its capability to generate diverse lexical items and context-rich reading tasks.
4. The teacher-made test was found to be more effective in evaluating grammatical knowledge.
5. Despite having difficulties, challenges, and inconsistencies in the grammar items of ChatGPT, it can still design well-structured, discriminating items that can align with student levels, especially in vocabulary and reading comprehension skills.
6. ChatGPT helps teachers generate test items with a lower cost, faster pace, and less human intervention.
7. The importance of the integration of AI tools into EFL assessment practices as a complementary resource rather than a replacement for teacher involvement.
8. Combining AI-generated assessments with teacher-designed tests enhances both efficiency and validity in language evaluation.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

9. The accuracy of the use of the prompts while asking for information affects the responses given by the chatbot.
10. When AI tools such as ChatGPT are integrated with human expertise carefully in language assessment contexts, they can work effectively, validly, and reliably.

References

Abdelhameed, A. A. (2020). Developing EFL Reading Comprehension Skills through a Suggested Program Based on a Communicative Language Teaching Approach among Second Year Faculty of Specific Education Students. *Journal of Faculty of Education* , 19-36.

(AERA), American Educational Research Association, (APA), American Psychological Association & (NCME), National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Akpan, M. (2025). Have we reached artificial general intelligence? Comparison of ChatGPT, Claude, and Gemini to human literacy and education benchmarks. Corporate Ownership & Control, 22(1), 103–110. https://doi.org/10.22495/cocv22i1art8

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

Alduais, A. M. (2013). *Language Testing.* Deutschland: LAP LAMBERT Academic Publishing.

Alkhateeb, A., Hezam, A. M. M., & Almuraikhi, A. A. (2025). Assessing the use of AI tools for EFL exam preparation at Saudi universities: efficiency, benefits, and challenges. *Cogent Education, 12*(1), 1-19. https://doi.org/10.1080/2331186x.2025.2507553

Alsagoafi, A. A., & Alomran, H. S. (2025). Revolutionizing Assessment: Leveraging ChatGPT for Automated Item Generation: An AI Driven Exploratory Study with EFL Teachers. *World Journal of English Language, 15*(6), 385-394. https://doi.org/10.5430/wjel.v15n6p385

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence, 6.*

Ayala-Pazmiño, M. F., & Alvarado-Lucas, K. I. (2023). Integrating Artificial Intelligence into English Language Education in Ecuador: A Pathway to Improved Learning Outcomes. *Digital Publisher CEIT, 8*(3), 679-687. doi:https://doi.org/10.33386/593dp.2023.3-1.1862

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford Universty Press.

Bailey, D. (2019). Chatbots as Conversational Agents in the Context of Language Learning. *The Fourth Industrial Revolution and Education*, 32-40.

Baxter, A. (1997). *Evaluating Your Students.* London: Richmond Pubishing.

Bekleyen, N. (2010). An Examination of Language Achievement Tests Administered in Primary Education. *Eurasian Journal of Educational Research*(41), 19-35.

Bieleke, M., Goetz, T., Krannich, M., Roos, A., & Yanagida, T. (2021). Starting Tests With Easy Versus Difficult Tasks: Effects on Appraisals and Emotions. *The Journal of Experimental Education*, 317-335. doi:https://doi.org/10.1080/00220973.2021.1947764

Brown., H. D. (2004). *Language Assesment: Principles and Classroom Practices.* USA: Pearson Education,Inc.

Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices.* Pearson Education.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., GirishSastry, Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., & Dan. (2020). Language Models are Few-Shot Learners. *34th Conference on Neural Information Processing Systems*, 1-25.

Carroll, J. B. (1973). Foreign Language Testing: Will the Persistent Problems Persist? *ERIC*, 1-20.

Chun, J. Y., & Barley, N. (2024). A Comparative analysis of multiple-choice questions: ChatGPT-generated items vs. human-developed items. In G. H. Carol A. Chapelle, *Paths for Exploring AI in Applied Linguistics* (pp. 118-136). Iowa State University Digital Press. doi:https://doi.org/10.31274/isudp.2024.154.01

Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory.* Cengage Learning.

Coşgun, G. E. (2025). Artificial intelligence literacy in assessment: Empowering pre-service teachers to design effective exam questions for language learning. *British Educational Research Journal (BERJ), 51*, 2340-2357. doi:https://doi.org/10.1002/berj.4177

Danesi, M. (2024). *AI in Foreign Language Learning and Teaching: Theory and Practice.* New York: Nova Science Publishers, Inc.

Dergaa, I., Chamari, K., Zmijewski, P., & Saad H. B. (2023). From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing. *Biology of Sport, 40*(2), 615-622. doi:https://doi.org/10.5114/biolsport.2023.125623

El Hassan F.A.M. and Alsalwah A.F. (2025). Exploring the Impact of ChatGPT on EFL Reading Practices: Opportunities and Challenges. *International Journal of English Language Teaching, 13*(1), 85-93. doi:https://doi.org/10.37745/ijelt.13/vol13n18593

Field, A. (2013). *Discovering Statstics Using IBM SPSS Statistics* (4th Edition ed.). London: SAGE Publications Ltd.

Fraenkel, J. R., Wallen N. E., & Hyun, H. H. (2012). *How to Design and Evaluate Research in Education* (8th Edition ed.). New York: The McGraw-Hill Companies, Inc.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment.* New York: Routledge.

Galaczi, E. & Luckin, R. (2024). *Generative AI and Language Education: Opportunities, Challenges and the Need for Critical Perspectives.* Cambridge University Press & Assessment .

George, D., & Mallery, P. (2016). *IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference* (14th Edition ed.). New York: Routledge.

Gehringer, E. F. (2004). Reuse of homework and test questions: when, why, and how to maintain security? *Frontiers in Education Conference*, 24-29.

Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice.* USA: Cambridge University Press.

Gutiérrez, L. M. (2023). Artificial Intelligence in Language Education: Navigating the Potential and Challenges of Chatbots and NLP. *Research Studies in English Language Teaching and Learning, 1*(3), 180–191. doi:https://doi.org/10.62583/rseltl.v1i3.44

Hadzhikoleva, S., Rachovski, T., Ivanov, I., Hadzhikolev, E., & Dimitrov, G. (2024). Automated Test Creation Using Large Language Models: APractical Application. *Applied Science, 14*, 1-19. doi:https://doi.org/10.3390/app14199125

Hamad, A., M. (2025). An Overview of Multivariate Statistical Methods and Their Practical Applications. *Jurnal Pendidikan Matematika, 3*(1), 16. https://doi.org/10.47134/ppm.v3i1.2084

Hansen, H. & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. *Proceedings of the 44th Annual Conference of the Cognitive Science, 44*(44). doi:https://escholarship.org/uc/item/44s454ng

Harunasari, S. Y. (2023). Examining the Effectiveness of AI-integrated Approach in EFL Writing: A Case of ChatGPT. *International Journal of Progressive Sciences and Technologies (IJPSAT), 39*(2), 357-368.

Hatem, G., Zeidan, J., Mathijs, G., & Moreira, C. (2022). Normality Testing Methods and the Importance of Skewness and Kurtosis in Statstical Analysis. *BAU Journal - Science and Technology, 3*(2). doi:https://doi.org/10.54729/KTPE9512

Heaton, J. (1990). *Classroom Testing.* NewYork: Longman Group UK Limited .

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence In Education.* Boston: The Center for Curriculum Redesign.

Hsiao, Y., Klijn, N., & Chiu, M. (2023). Developing a Framework to Re-design Writing Assignment Assessment for the Era of Large Language Models. *Learning: Research and Practice*, *9*(2), 148–158. doi:https://doi.org/10.1080/23735082.2023.2257234

Hughes(a), A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Hughes(b), A. (2003). Testing Oral Ability. In A. Hughes, *Testing for Language Teachers* (Second Edition ed., pp. 113-135). Cambridge: Cambridge University Press.

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 15873–15892*, 15873–15892. doi:https://doi.org/10.1007/s10639-023-11834-1

John W. Oller, J. (1979). *Language Test at School*. London: Longman Group Ltd.

Johnsson, N. (2023). AI driven Test Case Generation: An In-Depth study on the Utilization of Large Language Models for Test CaseGeneration.

Jur¯ane-Br¯emane, A. (2023). Digital Assessment in Technology-Enriched Education: Thematic Review. *Education Science, 13*, 1-13. doi:https://doi.org/10.3390/educsci13050522

Kanık, M. (2024). The use of ChatGPT in assessment. International Journal of Assessment Tools in Education, 11(3), 608-621. https://doi.org/10.21449/ijate.1379647

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 1-13. https://doi.org/10.1016/j.lindif.2023.102274

Khin, N. N., & Soe, K. M. (2020). University Chatbot using Artificial Intelligence Markup Language. *2020: IEEE Conference on Computer Applications(ICCA)*. doi:http://dx.doi.org/10.1109/ICCA49400.2020.9022814

Koç, F. Ş., & Savaş, P. (2025). The Use of Artificially Intelligent Chatbots in English Language Learning: A Systematic Meta-Synthesis Study of Articles Published between 2010 and 2024. *ReCALL, 37*(1), 4-21. https://doi.org/10.1017/S0958344024000168

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, 1–14. doi: DOI: 10.1177/00336882231162868

Koraishi, O. (2023). Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment. *Language Education & Technology(LET Journal), 3*(1), 55-72.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2019). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Li, Y., Zhou, X., Yin, H., & Chiu, T. K. F. (2025). Design language learning with artificial intelligence (AI) chatbots based on activity theory from a systematic review. *Smart Learning Environments, 12*(24), 1-23. doi:https://doi.org/10.1186/s40561-025-00379-0

Liu, F. (2010). Reading Abilities and Strategies: A Short Introduction. *International Education Studies, 3*(3), 153-157.

Luong, N. T., Linh, N. H., & Hanh, L. D. (2024). Teachers' Perspectives on AI-Driven Quillionz for Generating EFL Reading Comprehension Quizzes. *Proceedings of the AsiaCALL International Conference, 6*, 20-34. doi:DOI: 10.54855/paic.2462

Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT Literacy in Language Education: A Global Perspective and Comprehensive Approach. *Computers and Education Artificial Intelligence, 7*(1), 1-10. doi:http://dx.doi.org/10.1016/j.caeai.2024.100278

MacMillan, F. M. (2016). Assessing reading. In D. T. Banerjee, *Handbook of Second Language Assessment* (pp. 113-131). Boston/Berlin: Walter de Gruyter Inc.

Mbwambo, N. M., & Kaaya, P. B. (2024). ChatGPT in Education: Applications, Concerns and Recommendations. *Journal of ICT Systems*, *2*(1), 107–124. doi:DOI: 10.56279/jicts.v2i1.87

Messick, S. (1987). *Validity*. New Jersy: Educational Testing Service.

Mengual-Macenlle, N., Marcos, P. J., Golpe, R., & González-Rivas, D. (2015). Multivariate analysis in thoracic research. *Journal of thoracic disease*, *7*(3), E2–E6. https://doi.org/10.3978/j.issn.2072-1439.2015.01.43

Moon, H., Chung, Y., & Randolph, A. W. (2025). Teaching and Learning Languages with ChatGPT: Challenges and Opportunities in Multilingual Classrooms in Higher Education. *Indonesian Journal of English Language Teaching and Applied Linguistics, 10*(1), 207-223. doi:http://dx.doi.org/10.210

Munby, J. (1978). *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Navarro, D. (2019). Learning Statistics with R - A tutorial for Psychology Students and other Beginners (version 0.6). *Statistics LibreTexts. https*. Retrieved from https://stats.libretexts.org/Bookshelves/Applied_Statistics/Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)

Nunnally, J. C. (1978). *Psychometric Theory* . USA: McGraw-Hill, Inc.

O, K. M. (2024). A comparative study of AI-human-made and human-made test forms for a university TESOL theory course. *Language Testing in Asia*, *14*(19), 1-17. doi:https://doi.org/10.1186/s40468-024-00291-3

Olshtain, E. (2001). Functional Tasks for Mastering the Mechnics of Writing and Going just Beyond. In M. Celece-Murcia, *Teaching English as a Second or Foreign Language* (pp. 207-2017). United State of America: Heinle & Heinle, Thomson Learning.

Ridwan, R. N. (2024). The Influence of Vocabulary Mastery and Grammar on Students' Reading Comprehension Narrative texts. *Verba: Journal of Applied Linguistics, 3*(1), 41–50.

Rost, D. H. (1989). Reading comprehension: Skill or Skills? *Journal of Research in Reading, 12*(2), 87-113.

Säuberli, A., & Clematide, S. (2024). Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models. *3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, 22–37. https://aclanthology.org/2024.readi-1.3/

Sembey, R., Hoda, R., & Grundy, J. (2024). Emerging Technologies in Higher Education Assessment and Feedback Practices: A Systematic Literature Review. *The Journal of Systems & Softwar*, 1-18. doi:https://doi.org/10.1016/j.jss.2024.111988

Sihite, M. R., Meisuri, & Sibarani, B. (2023). Examining the Validity and Reliability of ChatGPT 3.5-Generated Reading Comprehension Questions for Academic Texts. *Randwick International of Education and Linguistics Science (RIELS) Journal, 4*(4), 937-944. doi:https://doi.org/10.47175/rielsj.v4i4.835

Shin, D. (2023). A Case Study on English Test Item Development Training for Secondary School Teachers Using AI Tools: Focusing on ChatGPT. *Language Research, 59*(1), 21–42. doi:https://doi.org/10.30961/lr.2023.59.1.21

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27-40. doi:https://hdl.handle.net/10125/73530

Shin, D., Kwon, K. S., & Lee, Y. (2025). Examining the efficacy of generative artificial intelligence in item generation: comparative analysis of human-developed and AI-generated reading tests. *Education and Information Technologies*. doi:https://doi.org/10.1007/s10639-025-13683-6

Singha, S., Singha, R., & Jasmine, E. (2024). Enhancing Language Teaching Materials Through Artificial Intelligence: Opportunities and Challenges. In F. Pan, *AI in Language Teaching, Learning, and Assessment* (pp. 22-42). IGI Global Scientific Publishing.

Steven M. Downing and Thomas M. Haladyna. (2011). *Hand Book of Test Development.* Taylor & Francis.

Stiggins, R. J. (1993). High Quality Classroom Assessment: What Does It Really Means? *ITEMS. Instructional Topics in Educational Measurement*, 35-39.

Stiggins, R. J., Arter, J. A., Chappuis, J., &Chappuis, S. (2004). *Classroom Assessment for Student Learning: Doing It Right – Using It Well.* Portland, Oregon: Assessment Training Institute, Inc.

Sugawara, S., Nangia, N., Warstadt, A., & Bowman S. R. (2022). What Makes Reading Comprehension Questions Difficult? *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 6951–6971. Retrieved from https://aclanthology.org/2022.acl-long.479.pdf

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction, 95*.

Swaie, M., & Algazo, M. (2023). Assessment purposes and methods used by EFL teachers in secondary schools in Jordan. *Front. Educ., 8*, 1-10. doi:https://doi.org/10.3389/feduc.2023.1192754

Thao, N. T. (2023). The Application of ChatGPT in Language Test Design –The What and How. *Proceedings of the AsiaCALL International Conference, 4*, 104-115. doi:https://doi.org/10.54855/paic.2348

Thaqi, E., Mantawy, M. & Kasneci, E. (2024). SARA: Smart AI Reading Assistant for Reading Comprehension. *ETRA*, 1-3. doi:https://doi.org/10.1145/3649902.3655661

Urquhart, A.H., & Weir, C.J. (1998). *Reading in a Second Language: Process, Product and Practice.* New York: Taylor & Francis.
Verhoeven, L., & Perfetti, C. (2008). Advances in Text Comprehension: Model, Process and Development. *Applied Cognitive Psychology, 22*(3), 293–301. doi:https://doi.org/10.1002/acp.1417

Wahyudin, A. Y., Aminatun, D., Mandasar, B., Hamzah, I., Ayu, M., Oktaviani, L., & Alamsyah, R. (2024). *Basic Principles of English Language Teaching.* Bandar Lampung: Universitas Teknokrat Indonesia.
Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand, 14*, 1-19. doi:https://doi.org/10.1016/j.amper.2025.100224

Weir, C. J. (1991). *Communicative Language Testing.* UK: Pearson ESL.

Weir, C. J. (2005). *Language Testing and Validation.* NewYork: PALGRAVE MACMILLAN.

Yaneva, V., & Davier, M. V. (2023). *Advancing Natural Language Processing in Educational Assessment.* New York: Routledge.

Yatri, D., Anugerahwati, M., & Setyowati, L. (2023). Artificial Intelligence (AI) in L anguage Learning (English and Arabic Class): Stdents' and Teachers' Experience and Perceptions. *Transformational Language Literature and Technology Overview in Learning (TRANSTOOL), 3*(1), 1-12. doi:DOI: 10.55047/transtool.v3i1.1338

Zafar, N., Saira and Afzal, S. (2025). AI-Powered Reading Support for Multilingual Learners in Higher Education: A Critical Review. *Journal for Social Science Archives, 3*(1), 776 – 786.

Zyska, D., Dycke, N., Buchmann, J., Kuznetsov, I., & Gurevych, I. (2023). CARE: Collaborative AI-Assisted Reading Environment.*ACL Anthology*. https://doi.org/10.18653/v1/2023.acl-demo.28

Appendices

Appendix 1: the test formats of ChatGPT- generated and teacher –made test formats

Reading Comprehensic Test
University of Halabja
Level: EFL – Second Year University Students College
of Basic Education
Duration: 2 Hours
Department of Language
Total Marks: 40 Mid-

2024-2025

Student name: ……..…

1. AI-generated test

Part A: Text Passage

Title: The Influence of Technology on Daily Life

In recent decades, technology has transformed nearly every aspect of our lives. From the way we communicate to how we work and learn; technological innovations have become deeply integrated into daily routines. For instance, smartphones allow instant access to information, and video conferencing makes it possible to attend global meetings without leaving home. However, these conveniences come with challenges.

A growing body of research suggests that excessive reliance on technology can impact mental health, especially among young adults. Many find it difficult to disconnect, resulting in increased anxiety and sleep disturbances. While social media platforms offer ways to stay connected, they also promote comparison, distraction, and sometimes misinformation.

Despite these concerns, the benefits of technology are undeniable. Online learning platforms, for example, provide flexible education for people in remote areas. Similarly, wearable health devices track vital signs and encourage physical activity. In healthcare, artificial intelligence is helping doctors make faster and more accurate diagnoses.

Still, one must ask: is society too dependent on technology? The key lies in balance—using digital tools thoughtfully without letting them control our lives. As we continue to integrate technology into our personal and professional environments, individuals must develop critical thinking skills to evaluate information and choose when, how, and why to engage with digital platforms.

 Section A: Vocabulary & Grammar **(6 Marks)**

1. Choose the best meaning for the underlined word from the text. (4 Marks).

a) “excessive”

- a. moderate
- b. too much
- c. helpful
- d. missing

b) “misinformation”

- a. false or misleading information

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

•b. official news
•c. shared opinions
•d. marketing material

2. Identify the grammatical function of the phrase: "resulting in increased anxiety" (2 Marks)
•a. Noun clause
•b. Prepositional phrase
•c. Participle phrase
•d. Gerund phrase

 Section B: Comprehension Skills **(14 Marks)**

3. Scanning & Skimming
•a) What is the main idea of the passage?
……
……
……
……
•b) List two examples the author uses to show the impact of technology**.**
.....
.....
.....

4. Activating Background Knowledge & Inference
- a) From your own experience, how has technology changed your education? **(Open-ended)**

……
……
……
……..
- b) What can you infer about the author's attitude toward technology?

……
……
……
……..

5. Guessing Meaning from Context
Guess the meaning of *"critical thinking"* as used in the last paragraph.
……
……
……
……..

6. Recognizing Purpose and Hidden Information
- a) What is the purpose of this text? (Choose one)

○a. To entertain
○b. To argue
○c. To inform and raise awareness
○d. To criticize technology

- b) Find one idea that is implied but not directly stated.

……
……
……..

7. Understanding Text Structure

- a) Identify a sentence that shows contrast.

.....
.....
.....

- b) How does the last paragraph connect with the introduction?

……
……
……
……..

---

2. Teacher-made test

Part B/…….

Mathilde Loisel was one of those pretty and charming girls born into a family of *artisans*, she married to a clerk in the Ministry of Education. She suffered endlessly, feeling herself born for every delicacy and luxury. She suffered from the poorness of her house, from its mean walls, worn chairs, and ugly curtains. She feels stuck in a life that believes it is far beneath her.

One day, her husband brings home an invitation to a fancy ball, hoping to make her happy. But instead of excitement, Mathilde feels *ashamed*—she has nothing elegant to wear. Her husband scrapes together money for a dress, and she borrows a sparkling diamond necklace from her rich friend, Madame Forestier. At the party, Mathilde feels amazing, finally living her dream for one magical night. But tragedy hits when she loses the necklace. They searched in the folds of her dress, in the folds of the coat, in the pockets, everywhere. They could not find it. Her husband borrows money to buy another necklace as a replacement.

They spend the next ten years working themselves into the ground to pay it off. Their lives are completely changed—they fall into deep poverty and hardship. Years later, Mathilde learns the shocking truth: the original necklace was fake and worthless.

Answer the following questions: (14 Marks)

1) What does Mathilde Loisel suffer from?

……
……

2) What is Mathilde and her husband's big shock?

……
……

3) What does she borrow from her friend?

……
……

4) Write a title for the story based on your understanding.

……
……

5) What is the main idea of the passage?

……
……

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6

……
……..
6) What do you understand about the feeling of Mathilde's husband after bringing the invitation? And why?
……
……
……..
Vocabulary section (4 marks)
Write the synonym of the following words from the passage.
1) Gather: …….. 2) Office worker: ……..
3) Dreams of wealth: ……. 4) Being very poor: …….
Grammar section
B: Correct the grammatical errors (3 Marks).
1) She happy but said she had nothing to try on.
……
……
2) They buy a new necklace which cost them a lot of money.
……
……

Wish you all the best

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 6