International Journal of Educational Innovations
ISSN 3078-5677

International Journal of Educational Innovations
Volume 2, Issue 1, 124-138
https://doi.org/10.46451/ijei.260110

Received: 5 May, 2025
Accepted: 5 September, 2025
Published: 1 January, 2026

Artificial Intelligence as a Reading Comprehension Test Designer

Nare Hakobyan
V. Brusov State University, Armenia
(Email: narekaytser@gmail.com)

Abstract
Despite their popularity, the design of reading comprehension tests in the form of multiple-choice questions (MCQs) is a time-consuming and arduous endeavor. For these reasons, quite often EFL teachers avoid the design of new tests for supplementary reading materials and the redesign of less useful ones. In this context, the use of Artificial Intelligence (AI) is valuable, more specifically an AI application Questgen that can generate reading comprehension tests within seconds, regardless of the text length or content. However, there is a shortage of data on the adequacy of such platforms and the impact they have on students' performance. The aim of the current paper is to identify how AI performs as an MCQ generator to check reading comprehension compared to teacher-designed tests, by pointing out its strengths and weaknesses. This mixed-method research includes 10 reading comprehension tests, weekly introspective questionnaires, teacher interviews, and a focus group discussion as data collection tools. The research questions included in the study are: 1) How are AI-generated tests different from teacher-designed tests?, 2) How do students perform dependent on the tests?, and 3) What are teachers' and students' attitudes to AI-generated vs. teacher-designed tests?

Keywords
Artificial Intelligence (AI), reading comprehension tests, MCQs, AI-generated tests, attitudes

Introduction

Artificial Intelligence (AI) is a branch of computer science that aims at perfecting and automating computational processes to the extent that they imitate human mental capacities. Even though AI has been actively developing since the 1960s, notable progress has been made since the COVID-19 pandemic. Naturally, there was an intense need to look for alternatives that would help humans adapt to the new reality with its existing challenges. This, of course, included TEFL (Teaching English as a Foreign Language).

In the context of TEFL, the pandemic was a time when teachers experienced a lack of teaching resources while transitioning to online teaching. In parallel, they were also lacking the skills to make an adequate transition and to design the necessary instructional materials. However, a major issue is that even though many AI tools and apps are available online, not all of them have been tested in terms of their adequacy (Golonka et al., 2014; Jiang, 2022). The majority of publications have been computer science-oriented, where the primary focus is on app implementation with no specific focus on how those apps performed in terms of language generation and what impact they had on learning. Unfortunately, not much has been done about measuring the effectiveness of different AI platforms in the process of language learning with TEFL-oriented results and their analysis. Even though AI has positively impacted most EFL skills (Bailey et al., 2020; Ebadi & Ebadijalal, 2020; Yong, 2020), there has been a limited amount of research on the effectiveness of particular AI tools in TEFL (Dodigovic, 2013; Dodigovic & Tovmasyan, 2021; Sharadgah & Sa'di, 2021), more specifically in reading comprehension test designs.

Reading is one of the most used skills necessary for staying informed, extensive reading, digital communication with friends and relatives, studying, and learning (van den Broek & Kendeou, 2022). Reading maintains its importance in English as a Foreign Language (EFL) as well, particularly in the early stages of EFL learning. In this sense, AI is invaluable for the design of reading comprehension tests to check reading comprehension and progress in an EFL classroom and in autonomous learning. Moreover, the creators of standardized EFL tests constantly seek alternatives to design new reading comprehension tests, which is challenging both financially and in terms of resources. This paper aims at filling in the existing gaps on how useful the AI platform Questgen is for designing reading comprehension tests in comparison to teacher-designed tests, how students perform depending on the tests, as well as what attitudes both students and teachers hold to AI.

Literature Review

MCQs as an assessment technique in reading comprehension

Reading comprehension is “an asynchronous, written two-way interaction between the author and reader, in which the reader’s primary task is to comprehend the author’s intent” (Cox et al., 2019, p. 118). Comprehension is interpreted as the “reduction of uncertainty” (Smith, 1971; 1973a, as cited in Sadeghi, 2021, p. 15) so that the match between whatever has been encoded by the writer is understood by the reader without “confusion” (Sadeghi, 2021, p. 16). Assessing reading comprehension is complex as it depends on several variables such as the text, its organization, type, vocabulary, the reader and their reading strategies, L1, aptitude, as well as context (Alderson, 2000; Brunfaut, 2022; Green, 2021; Sadeghi, 2021).

Moreover, checking reading comprehension as a receptive skill is not straightforward (Hughes, 2003). Therefore, test designers, who are mostly teachers, have to think about the assessment techniques for this skill (Hughes, 2003). So far, one of the most popular techniques has been multiple-choice questions (Hughes, 2003; Nation & Macalister, 2021), as it generated the most conventional and practical items applied in local and standardized language testing (Brown & Abeywickrama, 2018). They can check both lower-order or higher-order skills (Grabe & Yamashita, 2022; MacMillan, 2022). The former concerns checking whatever is explicitly written in the text, be it vocabulary or an idea, while the latter concerns the ability to read between the lines such as making inferences, identifying relations, summarizing, or evaluating. MCQs, which are comprehension checking questions, are usually “local” rather than “general” because they mainly focus attention on the “message of a particular text” (Nation & Macalister, 2021, p. 34) even though at times they can refer to more general ideas within a text.

MCQs possess a number of positive traits since they are dependable, fast in design, economical and familiar, but they also present some drawbacks as they are guessable, hard to design, and easily cheated (Bailey, 1998; Brunfaut, 2022). Among the most common errors in the design of these question types are grammar inconsistencies, syntactic inconsistencies (called "impure items"), having no correct or multiple correct answers, including prompts that give away the key (called "extraneous cues"), or having structural discrepancies among the options that lead to 2 vs. 2 or 3 for 1 grouping (Coombe et al., 2007). Successful MCQs must have similar items both linguistically and structurally to reduce guessing risks. They must also be cognitively or linguistically the least difficult to assess the comprehension of the text and not of the options (Brunfaut, 2022) by including the necessary information to find the key as the correct answer and exclude distracters as the wrong ones, by avoiding double negatives and by using familiar vocabulary in the item (Green, 2021). Other factors impacting item difficulty are the type of information, which can be both abstract and concrete, and the type of match, which concerns identifying one piece of information or multiple pieces of information, inferences, and categorization of information.

The next factor is plausibility of distracting information which concerns the number of distracters and how much common information they share with the text (MacMillan, 2016). "Lexical cohesion" (MacMillan, 2016, p. 120) is another relevant factor which refers to the cohesion between the vocabulary in the item and the text. It refers to lexical repetition, synonymy, antonymy, superordinate repetition, hyponymic repetition, co-reference, labeling, and substitution (MacMillan, 2007, as cited in MacMillan, 2016, p. 120). Simple repetition is manifested via repeating words and phrases verbatim while complex repetition shares a common morpheme but uses a different part of speech. Simple synonyms are those with the same meaning and part of speech while complex synonyms have the same meaning but different word classes. Simple antonymy is expressed with the same word class which is not the case with complex synonyms. The next two types of lexical cohesion concern the relationship between lexical items in terms of generality and specificity (Hoey, 1991). Moreover, whenever the key shares more vocabulary with the text, it tends to be easier (Freddle & Kostin, 1993; Kirsch, 2001), and MacMillan (2016) found that the key would lexically be more cohesive than the distracters. All these factors were considered in the qualitative analysis of MCQs in the current research because they are central to the identification of the key differences between the tests.

AI and MCQs

The use of AI for generating MCQs to assess reading comprehension also warrants consideration. AI is defined as the ability of machines or technologies to think, i.e., to understand, to process, and to generate human language similar to human beings (Russel & Norvig, 2010). Questgen, an AI app implemented in this research, is a generative AI application that uses Natural Language Processing (NLP) and generates questions of different types, e.g., MCQs, fill-in-the-blank questions, match-the-following, or true/false questions. Considering the popularity of MCQs, we focused on them.

Surprisingly, not much has been done to investigate the uses of AI in EFL assessment (Jiang, 2022), and research carried out thus far has indicated limited efficacy (Golonka et al., 2014). According to the review of 64 articles published between 2015-2021 (Sharadgah & Sa'di, 2021), AI was deemed to have a positive impact on EFL assessment with automated feedback, translation (Bahri & Mahadi, 2016), writing (Gao, 2021; Hamuddin et al., 2020; Wang, 2022; Yong, 2020), and speaking (Ebadi & Ebadijalal, 2020; El Shazly, 2021). Additionally, AI helps to improve students' vocabulary (Lai & Chen, 2021; Tai et al., 2022) and grammar (Abu Ghali

et al., 2018; Kim, 2019) by fostering students' motivation (Aljohani, 2021; Al-mawaly & Al-Jamal, 2022). However, no research has been conducted with the specific focus on how AI-generated tests are different from teacher-designed tests and how students' performance is impacted by those differences in the EFL context.

Some assessment-related research was conducted in the medical context where the researchers examined how professorial staff generated MCQs based on the medical coursebooks vs. how AI generated them (Cheung et al.,j 2023). The findings showed that AI was faster in the test generation and significantly less relevant because it would generate options not discussed content-wise in the input. The researchers claimed that AI has the potential to generate comparable quality MCQs for medical examinations. In the research by Law et al. (2025) similar findings were acquired for the medical PEEM examination: AI-generated tests were easier and generated faster but were inaccurate and targeted low-order skills. In another study, the adoption of different AI tools for EFL exam preparation was examined through a teachers' survey. The results showed that MCQs were among the most dominating techniques that the AI tools had generated though they featured "inconsistent, inaccurate, and irrelevant information" (Alkhateeb et al., 2025, p. 13).

AI-generated tests have also been examined in industrial robotics education. Among the weaknesses observed in the tests were repetitive question structures which would lead to inconsistencies, irrelevance of the content, and oversimplified distracters (Pöysäri et al., 2025). Despite quite similar results, the studies mentioned did not examine the linguistic features holistically, which are the reasons for the aforementioned weaknesses. Moreover, the tests in the above studies checked the discipline-related knowledge of the students which may not be acquired through the reading only. Therefore, the question of how EFL reading comprehension is assessed through AI-generated MCQs from a linguistic perspective remains unresolved and is addressed in the present study. Hence, this study investigates the following research questions:
1) How are AI-generated tests different from teacher-designed tests?
2) How do students perform dependent on the tests?
3) What are teachers' and students' attitudes to AI-generated vs. teacher-designed tests?

Methodology

Participants, procedure and materials

The aim of the current paper is to identify how AI performs as an MCQ generator to check reading comprehension compared to teacher-designed tests by detecting its strengths and weaknesses. The research included 48 high-school students aged 13-16 that had reached an intermediate proficiency level at school. They were native Armenians who studied English as a foreign language and were engaged voluntarily. The teachers had received consent for participation from the learners' parents prior to the research. The teachers were native-Armenian speakers with more than 4 years of EFL teaching experience. All of them had graduated from an American university with a required course in English Language Assessment.

This was a six-week quasi-experimental research with mixed methods. Five groups consisting of 9-10 students were exposed to the same weekly tests, each consisting of a teacher-designed test and an Artificial-Intelligence-generated test with an introspective questionnaire about the tests. The tests were presented in a random order each week. The students had no idea which test was designed by AI or by the teacher. In the sixth week the students participated in a focus group discussion and the teachers were interviewed after sharing all of the tests with them.

The texts were profiled to match the 95-98 vocabulary coverage (Nation, 2013) for Intermediate proficiency level students. They were expository about stress, reasons why teenagers get tattoos, floods, and the behavior of a good employee, and narrative, such as the story of a house renovation because of an incident with carrots.

Data collection instruments

Teacher-designed vs. AI-generated tests

The research data were collected with five multiple-choice reading comprehension tests each including one teacher-designed (T-D) and one AI-generated (AI-G) test. The first test consisted of 4 T-D and 4 AI-G questions, the second test 7 T-D and 5 AI-G questions, the third test 5 T-D and 4 AI-G questions, the fourth test 5 T-D and 8 AI-G, and the fifth test 4 T-D and 10 AI-G questions.

Introspective questionnaire

The same questionnaire with 5 reflective questions was applied after each testing session to gain understanding of test differences and similarities for each week's tests. The questions were about the tests' differences, difficulty, improvement, better comprehension check, and their favorite test.

Interviews

The interview consisted of three parts each containing 5 questions with a 10-minute break between the sessions. The participants were asked about the test differences and similarities, weaknesses and strengths, as well as their opinion about AI and its future application in their teaching.

Focus group discussion

This discussion included questions about students' general opinions on the integration of AI into testing. Twenty-five students were asked about its possible advantages, disadvantages, and differences compared to T-D tests. At this point, it was revealed to the participants which test was T-D and which one was AI-G.

Data analysis

The first question was answered with the help of a comparative analysis of both stems and options, as well as the vocabulary share measured with LexTutor. Moreover, the surveys and post-interview answers were analyzed via thematic coding. To answer the second research question, the internal consistency of both AI-G and T-D tests was measured. In addition, an Independent-sample t-test and ANOVA were conducted. The last question was answered with the quantified questionnaire replies, post-interview answers, as well as focus group results.

Results

How are AI-G tests different from T-D tests?

In order to answer this research question, weekly questionnaires, teacher interviews and comparative analysis of both stems and options were applied. Table 1 details the weekly test results and the questionnaire responses for each week in percentages about the differences between the tests, comparative difficulty, the best checking of comprehension, and possible improvements.

 ISSN 3078-5677

Table 1
Test Results and Introspective Survey Responses

Highest mean	Difference	Difficulty	Improvement	Comprehension
T-D (3.542)	Difficulty (25%), Content of the questions (20.9%) Length of items (16.7%)	T-D (37.5%)	AI-G (35.4%) None (33.3%)	T-D (60.4%)
T-D (4.792)	Test difficulty (22.9%), Paragraphs mentioned (16.7%) Content focus (10.42%).	T-D (41.7%)	AI-G (35.42%)	T-D (43.75%)
AI-G (2.958)	Test difficulty (25%), the Number of details (12.5%) Number of questions (4.17%)	T-D (50%)	None (31.25%) AI-G (25%)	T-D (43.75%)
AI-G (5.583)	Question numbers (22.92%) Test difficulty (18.75%) Questions with paragraphs (12.5%)	T-D (41.67%)	None (45.83%) T-D (29.17%)	AI-G (56.25%)
AI-G (7.583)	Difficulty (45.83%) Question number (14.58%) Details (14.58%)	AI-G (39.58%)	None (39.58%) T-D (29.17%)	AI-G (64.58%)

According to the teachers, the most notable differences were the number of the questions (25%), similar options in terms of content (25%), and longer questions and options in T-D tests (25%). The teachers thought that AI-G tests had the pros of being detailed enough (25%), comprehensible (25%), and having similar distracters (25%), as well as being comparatively easy (25%). Among the cons were too simple questions and distracters (75%), repetitive questions (13%), and missing keys (13%).

As for pros in T-D tests, the teachers mentioned that their items required critical thinking (25%), were designed based on their students' preferences for tests (25%), included different types of questions (25%) such as open-ended questions (60%), statements that had to be completed with the key (25%), negative questions (16%), or general questions (4%). The teachers also mentioned some cons in their tests such as easy questions with predictable answers (25%), not homogeneous options (25%), long questions with unknown words (13%), and fewer questions (13%) with less options (13%).

As for the comparative analysis, firstly, the tests were compared in terms of shared vocabulary outlined in Table 2. As shown, the T-D tests for weeks 1 and 2 had more shared vocabulary with the text than the AI-G tests. For the third week both tests were fairly close in their shared vocabulary with the text. For the last two tests it was the AI-G tests that shared comparatively more vocabulary than the T-D tests with the input. The average number of words was 8.8 in T-D items while it was 3.9 in AI-G tests.

 ISSN 3078-5677

Table 2
Vocabulary Analysis with LexTutor

Tests	AI1	T1	AI2	T2	AI3	T3	AI4	T4	AI5	T5
Token Recycling Index (%)	47.1	54.5	54.5	61.6	49.8	49.8	65.1	50	47.1	40
Coverage 1st(s) in 2nd (%)	49.0	58.6	53.2	66.6	53.7	54.7	63.8	50	57.7	49

Next, the items were compared in terms of information type and form in the stem, the consistency and errors of the options, lexically cohesive features of the items with the text as well as the relevance to the question. During the first test, the AI-G test had only 4 detail questions (all of which were special questions about a detail) while T-D test had 3 detail questions (2 were special questions with "true" and "not true" and one about a detail) and 1 main idea special question. Four options were taken directly from the test in the T-D test while there were 5 in the AI-G test. The T-D test was very homogeneous. However, the last two questions had rather long options in the form of complete sentences.

In the AI-G test, there were two questions with inconsistent options by being both unparallel and 2 vs. 2: question 2 had two nominal options and two clauses. Interestingly, the two nominal options were not relevant to the question as a response because the question asked about the effect of stress ("(Q2) How does stress affect the respiratory system?") and the response certainly could not be "Hyperventilation". In question 3, the first two options were countable nouns in a plural form, the third option was a gerund and the last one was an uncountable noun with a participle as an attribute. In this case, they were unparallel and impure options. In this test, 4 keys were lexical repetitions, 1 distracter was a repetition, and the others were not mentioned in the text.

In the T-D test, 4 options were verbatim lexical repetitions from the text and one was "Not Given". One option was constructed with the vocabulary of the text as a distracting idea and one by nominalizing the verb in question 1. In the second question, options 2 and 4 were newly formulated sentences with lexical repetitions from the text. The third option was the opposite of the information mentioned in the text with the simple antonym ("too slowly" - "too fast"). In question 3, one option was a new idea not mentioned in the text while option 3 was a newly-formulated sentence with a simple antonym ("not in danger" - "at risk"). In the last question, option 1 was a sentence with a simple antonym ("minor issue" for "serious problem"). The key was a newly-formulated sentence with repetition from the text. The last two options were made-up sentences with one repetition from the text ("long-term") and hyponymic repetition ("feelings" - "sad" and "happy emotions").

For test 2, the T-D test had 6 detail-oriented (4 were negative) and one main idea questions all in the form of special questions. Six out of 7 questions had the paragraphs mentioned. In the same test, 11 options were taken directly from the text with one distracter expressed with a simple antonym ("possible" - "impossible"), 8 were paraphrased ideas from the text and 9 were not mentioned in the text. For the AI-G test, there were 11 distracters not mentioned in the text, 5 keys taken directly from the text and 2 paraphrased distracters expressed with a simple synonymy ("self-expression" - "personal expression") and complex synonymy ("fashion models" - "fashionable"). In terms of inconsistencies, in the second T-D test, the only problem noticed was the length of the options specifically in questions 2, 3, 4 and 6. For the AI-G test, there were a number of inconsistencies such as unparallel options and 2 vs. 2 in questions 1 and 2.

For test 3, in the AI-G test 12 distracters were not mentioned in the text at all except for question 2 which had one word while the rest were not included in the text such as “Urban development, large dams, and heavy rainfall” where only “dams” was from the text. “Fertile soil, fast-melting ice caps, and sunny weather” similarly borrowed only “weather” from the text while “sunny” was replaced with a simple synonym “warm”. As for the keys, they were borrowed verbatim from the text, in one case by removing some part from the original context such as “Deep snow combined with heavy rain and sudden warmer weather” as a key extracted from “Although deep snow alone rarely causes floods, when it occurs together with heavy rain and sudden warmer weather it can lead to serious flooding”. All of them were detail questions in a special question form.

In the T-D test, 6 distracters were not mentioned in the text at all, 11 were paraphrased keys and distracters, 1 was “Not Given” and 2 were directly taken distracters, one of which was the opposite in its meaning to the key (“Frozen ground” instead of “Hot weather and rain”) and the other which was a homograph to the key to confuse the students. Four of them were detail questions and one was a main idea question. Two details were inquired in the form of incomplete statements, one detail was expressed with “Which of the following is true?” and 2 more were special questions.

In the fourth test, the AI-G test had 15 distracters that were not mentioned in the text, 6 keys were taken directly from the text; 1 was a shortened key and another was a key that was neither discussed nor could be inferred as a correct answer from the passage because the text did not provide such information. Four distracters had been paraphrased such as “brag about what a fine job you did” was transformed into “(3) Take full credit for the success”. There were 5 distracters directly taken from the text but which were the opposite answers to the questions. Seven were detail questions in the form of special questions and one was a main idea question in a special question form. In the T-D test, 8 distracters were not mentioned in the text, 6 were taken from the text (4 were keys and 2 were distracters) and 6 were paraphrased options. Only 2 were a special question, 1 was a general question and 2 were incomplete statements. Four were detail questions and 1 was an inference question. All the questions had “According to the passage” and were placed after the paragraphs they referred to. However, to match the layout of conventional MCQs, the questions were placed after the text, hence the passage numbers lost their functions as there were more than 3 passages. Regarding inconsistencies, there was a typo in question 1 where option 4 was presented with a small letter in contrast to the rest. In question 2, one option was an adverbial modifier of place, the rest were objects. In question 5 in week 4, two options were with “to” and 2 were without. In terms of missing correct answers, “d” could not be the key for question 3 in the T-D test as the text contained the opposing information to the key. For the AI-G test, the major inconsistency was unparallel and impure options in question 8.

In the last test, the AI-G test had 10 questions, all of which were detail questions in a special question form. Ten keys were taken directly from the text and only 1 key was paraphrased by changing “rinsed” to “washed”. Twenty-one distracters were not mentioned or discussed in the text at all while 8 were directly taken from the text. The T-D test had 4 detail questions in the form of special questions and with three options. One out of 4 was a negative question. Three of them had paraphrased distracters and keys in the form of full sentences. Special attention should be drawn to the absence of correct answers. In the AI-G test, question 3 was about the manner of how the carrots were cleaned while the answers were neither with the prepositional constructions “by” or “with” nor in a passive voice. Moreover, the first item in the T-D test did not have an actual correct answer because of the word choice in the key (“floor” was used

in the sense of "ground"). As noticed, the second predicate in the same sentence was in an incorrect tense.

How did the students perform dependent on the tests?

The internal consistency of each test was measured to answer the second research question. Cronbach's alpha for both tests was quite close and moderate as shown in Table 3.

Table 3

The Internal Consistency between AI-Generated And Teacher-Designed Tests

Test type	Chronbach's alpha	N of Items
Teacher-designed tests	.65	5
AI-generated tests	.60	5

As for the weekly performance, an Independent t-test was conducted to see how the groups performed against one another depending on the same weekly tests as detailed in Table 4. Then it was determined whether the test type (T-D or AI-G) was a factor of significant influence on groups' performance or not. According to the results of an ANOVA, for the first and third AI-G and T-D tests, the test type was not a factor of significant effect (F1 = 1.893, p1 = 0.172 and F3 = 1.432, p3 = 0.234). For the second, fourth and fifth tests, the test was a significant factor, impacting the students' performance (F2 = 11.268, p2 = 0.001, F4 = 27.926 and p4 < .001, F5 = 104.989 and p5 < .001).

Table 4

Weekly Independent Samples T-Test

						95% CI for Mean Difference	
	t	df	p	Mean Difference	SE Difference	Lower	Upper
Test 1	-1.376	94	0.172	-0.271	0.197	-0.662	0.120
Test 2	-3.357	94	0.001	[a] -1.604	0.478	-2.553	-0.655
Test 3	1.197	94	0.234	0.375	0.313	-0.247	0.997
Test 4	5.285	94	< .001	[a] 2.063	0.390	1.288	2.837
Test 5	10.246	94	< .001	[a] 4.688	0.457	3.779	5.596

Note. Student's t-test.

What are teachers' and students' attitudes to AI-G vs. T-D tests?

Out of five teachers, the average rating for the AI-G tests was 3.5 out of 5. Most were willing to integrate them into their lessons (75%) and all would like to incorporate AI-G tests (100%) although with caution and control to allow for necessary adjustments. They shared that the students would like AI-G tests (100%) because for students it would not matter who designed the test; they would do their best to score high. According to the introspective questionnaires, the students' favorite tests per week were the following: AI-G (56.25%) for week 1, T-D (37.5%) for week 2, AI-G (37.5%) for week 3, AI-G (54.17%) for week 4 and AI-G (58.33%)

for week 5. Nevertheless, during the focus group discussion 25 students expressed a drastically negative attitude to the use of AI in their learning (100%). They expressed some concerns about the acceptability of AI-G tests and were worried that AI did not recognize them well enough in contrast to their teachers, hence those tests could be too hard for them and would hurt their learning (100%).

Discussion

The students mostly achieved higher scores in the AI-G tests and lower ones in the T-D tests. Their performance was not significantly impacted by the test type during weeks 1 and 3 but was in weeks 2, 4 and 5. These outcomes were connected to the items, their content and form. Evidently, AI-G tests had both more questions (31 vs. 25) and options (124 vs. 96) than the T-D tests. The majority of the questions were detail-oriented (30 vs. 21) bound to the "nature of MCQs" (Nation & Macalister, 2021, p. 34). Notably, detail-oriented questions are considered concrete content (MacMillan, 2022) which requires cognitively lower-order skills to answer them (Grabe & Yamashita, 2022). These questions demand explicit answers (Grabe & Yamashita, 2022; MacMillan, 2016) and local reading (Nation & Macalister, 2021, p. 34). In addition, T-D items contained more questions about the main ideas of the texts (4 vs. 2) and incorporated more paraphrased options both form- and content-wise (31 vs. 8). The paraphrasing was in the form of simple, complex antonyms, and synonyms, as well as homographs. Moreover, both tests verbatim borrowed words, phrases, and sentences from the texts with AI-G tests having more overlaps in keys (28 keys and 14 distracters) and T-D tests having more in distracters (12 keys and 19 distracters). Even though both tests had distracters not mentioned in the text content-wise, AI-G tests were approximately three times higher in this respect (AI-G tests - 70 and T-D - 24). As the figures show, AI shared less content with the text by having more distracters not mentioned in the text with the keys mainly borrowed from the input. For the students, it was easier to find the correct answers as the keys had been directly borrowed from the texts with the distracters not being included in the text at all; they were from Wordnet which suggested co-hyponyms that were correct but irrelevant to the text. In contrast, T-D tests shared more vocabulary overlap with the distracters which was done to make the process of finding the key harder and more confusing. Paraphrased vocabulary in the distracters could have been another contributing factor to such results. The students could have been confused more easily because of more elaborate distracters such as the ones expressed with homographs, synonyms, or antonyms. The distracters seemed more truthful and correct with the feeling of being discussed in the text. Having irrelevant and simplistic options and keys generated by AI was in keeping with previous studies (Alkhateeb et al., 2025; Cheung et al., 2023; Law et al., 2025; Pöysäri et al., 2025), however, what was revealed with the present research is that T-D tests also contained irrelevant options and keys, in some cases being cognitively too marked.

As for shared vocabulary, better results were observed in tests which shared more vocabulary. These findings are consistent with the claim that more shared vocabulary makes the test easier (Freddle & Kostin, 1993; Kirsch, 2001). Our results also support the idea that keys supposedly share more vocabulary with the text (MacMillan, 2016) as was the case in AI-G tests (28 keys) in contrast to T-D ones (12 keys).

In terms of inconsistencies, it is known that cognitively and structurally homogeneous options reduce the risk of guessing (Brunfaut, 2022). In the AI-G tests there were 10 inconsistencies and only 4 in the T-D tests. As for missing answers, AI-G tests had more of them (3 vs. 2). One key could not be inferred from the text and 2 were irrelevant form-wise in AI-G tests whereas one could not be inferred and one had a wrong word choice in the T-D tests. Because of these

inconsistencies, AI-G items could be answered easily as was the case with the only key in the past form in question 3 while the rest of the options were in the form of bare infinitives, which also occurred in the AI-G test in week 5. Overall, in AI-G tests 2 answers were missing due to incorrect form and 1 because of content, while both were missing content-wise in T-D tests. Additionally, the T-D tests did not follow another principle in MCQs which is not "including double negatives" (Green, 2021, p. 117) because those features make the item not about the text comprehension but about its own comprehension (5 double negative questions). However, what was surprising was that the T-D test with the most double negatives (4 out of 7 questions) led to higher scores in week 2 (mean = 4.792). This might be related to the shared vocabulary and the numbers of the paragraphs mentioned in the stem, which may have helped students' performance.

As for the length of the options, they were twice as long in the T-D tests; however, this might not lead to a positive effect if "simpler language than the text" (Green, 2021, p.117) was not applied in the items such as "infer" (week 4), "forecast" (week 3), "shallow" (week 3) in T-D tests. Less wording could have made the test less contaminated with reading and required less mental effort for decoding. However, this was not the case in week 1 when the students mentioned that AI-G test needed improvements by making both the options and questions longer (18.75%) because having short questions and one-word options made it hard to choose the correct option (6.25%) as they were not "long enough for detailed comprehension" (Green, 2021, p. 116) to identify the context or understand both the question and the options. Overall, the tests (T-D test in week 2, AI-G tests in weeks 4 and 5) with longer wording and familiar vocabulary led to higher scores.

For the results in week 4, the higher score in the AI-G test could be related to the distortion of the initial layout of the items which were removed from the text and placed after it. For the fifth week, even though the AI-G test had double the questions, the students scored high on that test. This was not unexpected because questions 1 and 4 were about the amount of money spent on carrots and drain cleaner with the distracters not being mentioned in the text at all. Question 3 had the extraneous clue of the past tense to the question with the auxiliary in the past whereas the distracters were in a bare infinitive form. In addition, the distracters were not mentioned in the text either. Despite not having a correct answer in T-D question 1 ("floor" vs. "ground"), it did not have a negative impact on the answers reasoned with the possibly interchangeable use of "floor" and "ground" between non-native English teachers and students. Questions 3 and 4 were elaborate in terms of their distracters which focused on "multiple pieces of information" (MacMillan, 2016, p. 117) rather than one which impacted the students' performance.

For the first and third tests, the students' performance did not significantly differ because of the tests. However, for the first week the mean was higher on the T-D test (mean = 3.542) which shared more vocabulary with the text and contained familiar vocabulary in the distracters, despite having a few paraphrased options. In contrast, the AI-G test shared less vocabulary with the text, with the majority of the distracters not being mentioned in the text. Moreover, those distracters were medical terms mostly unknown to the students, hence "more difficult" (25%) and needing "improvements" (35.4%), according to the students. For the third week, the mean was higher on the AI-G test (2.958). Similar to the first-week test, the distracters were not mentioned in the text, however, the vocabulary was familiar to the students. The T-D test contained unfamiliar words such as "forecast", "shallow", one homograph "unleashed" ("unlashed" in the text) as well as one main idea question about the whole text. Moreover, the distracters were elaborated quite carefully in terms of both content and vocabulary. That is why

the students marked it as a more difficult test (50% with the rest of the answers being distributed within "AI-G tests" and "None") that checked their comprehension of the text the best (43.75%).

According to the students, their favorite test was the AI-G test because it required "short and concrete answers" (6.25% for week 1), and was "easy to understand with direct answers" (25% for week 3, 22.92% for week 4 and 17.78% for week 5). For the second week, the students liked the T-D test because it "was easy" (8.3%). In essence, the students favored the AI-G tests with no awareness that they had been generated by AI. This was expected considering how widely it is applied on social networking sites and media platforms which have their impact on users, specifically on digital natives (Gee & Hayes, 2011). However, what is surprising is the drastic change of attitudes after revealing which tests had been generated by AI and their teachers in the focus group discussion. This might be connected to the bias related to AI locally such as "not thinking" and "cognitively not advanced". Additionally, as something new and unclear, there might be some resistance and defensiveness to its usage. So, students' attitudes were natural if the conservatism of the students' nationality is also taken into consideration.

What is more surprising was the teachers' positive attitude to AI integration, similar to the findings by Al-mawaly and Al-Jamal (2022) and Aljohani (2021). It was valued as extra assistance whenever they were short of time and needed something urgent. However, they did not express complete trust in AI and indicated the need to provide extra control over the tests to make sure that they were mistake-free and dependable. They were confident that the students would not mind using AI in their learning. Therefore, it is hard to imagine how AI is going to be integrated into EFL teaching considering the controversies over attitudes. Should the teachers use AI secretly? And what about students' trust?

Conclusion

In conclusion, even though T-D tests were slightly more consistent than AI-G tests, it can be recommenced that AI-G tests be integrated into reading comprehension test design. AI-G tests fulfilled their major task, which was to check students' text comprehension. For that, AI generated more questions and more options by focusing on almost each paragraph of the texts. Considering the amount of time spent on AI-G tests and the closeness in their internal consistency to the T-D tests, AI-G tests can become a huge asset for EFL teachers, autonomous learners, and standardized tests. In contrast, teachers came up with cognitively harder questions and options which were connected to the goal of the teachers to distinguish between their strongest, weakest, and middle-proficiency level students. This differentiation matters to provide an objective assessment, hence a trustworthy learning environment. Overall, teachers' control is recommended for AI-G tests because AI focuses on overall reading comprehension with no intention to identify the true reading proficiency of their students. Moreover, the opposing attitudes have to be considered as well because if students have negative attitudes towards AI usage, the pros become somewhat devalued. Based on the findings in this paper, teachers are warned about the possible attitudes and can start prior guidance before the actual application of AI.

Limitations

One of the limitations in the current research was the relatively small sample size. Another limitation was the students' age, which may have led to less authentic responses. Teachers' busy schedule was another limitation as it could have led to issues in the quality of the tests.

Recommendation

Future research would benefit from analysing students' performance after EFL teachers improve the AI-G tests, instead of using them in their original form. It would also be interesting to add the use of reading strategies as an extra variable in the research design. Moreover, tests generated by different AI platforms would be interesting to compare.

References

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

Abu Ghali, M., Abu Ayyad, A., Abu-Naser, S., & Abu Laban, M. (2018). An Intelligent Tutoring System for Teaching English Grammar. *International Journal of Academic Engineering Research*, 2(2),1-6.

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

Aljohani, R. A. (2021). Teachers and students' perceptions on the impact of artificial Intelligence on English language learning in Saudi Arabia. *Journal of Applied Linguistics and Language Research*, *8* (1), 36-47.

Al-mawaly, H. M., & AL-Jamal, D. A. (2022). The Effect Of Artificial Intelligence Application on Jordanian EFL Sixth-Grade Students' Listening Comprehension And Their Attitudes Towards It. *Journal of Positive School Psychology 6* (6), 8781-8791.

Bahri, H., and Mahadi, T. S. T. (2016). Google translate as a supplementary tool for learning malay: a case study at universiti sains malaysia. *Adv. Lang. Literary Stud.* 7, 161–167. http://dx.doi.org/10.7575/aiac.alls.v.7n.3p.161

Bailey, K. M. (1998). *Learning About Language Assessment: Dilemmas, Decisions and Directions.*

Bailey, D., Southam, A., & Costley, J. (2021). Digital storytelling with chatbots: mapping L2 participation and perception patterns. *Interactive Technology and Smart Education*, 18(1), pp.85-103. https://doi.org/10.1108/ITSE-08-2020-0170

Brown, D. H. & Abeywickrama, P. (2018). *Language Assessment: Principles and Classroom Practices (3rd Edition)*. Pearson.

Brunfaut, T. (2022). Assessing reading. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing (Second edition)*, 254-267. Routledge.

Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. (2023) ChatGPT versus human in generating medical graduate exam multiple choice questions - A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS ONE* 18(8): e0290691, 1-12. https://doi.org/10.1371/journal.pone.0290691

Coombe, C., Folse, K. & Hubley, N. (2007) *A practical guide to assessing English language learners*. Ann Arbor: The University of Michigan Press.

Cox, T. L., Brown, J., & Bell, T. R. (2019). In advanced L2 reading proficiency assessments, should the question language be in the L1 or the L2?: Does it make a difference? In P.Winke, S. M. Gass (Eds.), *Foreign language proficiency in higher education*, 117–136. Switzerland: Springer Nature. https://doi.org/10.1007/978-3-030-01006-5_7

Dodigovic, M. (2013). Intelligent Sentence Writing Tutor: A System Development Cycle, *International Journal of Artificial Intelligence in Education*, 22, 141 – 160.

Dodigovic, M. & Tovmasyan, A. (2021). Automated writing evaluation: The accuracy of Grammarly's feedback on form. *International Journal of TESOL Studies 3* (2), 71-87. https://doi.org/10.46451/ijts.2021.06.06.

Ebadi, S. & Ebadijalal, M. (2020). The effect of google expeditions virtual reality on efl learners' willingness to communicate and oral proficiency. *Computer-assisted Language Learning* 33, pp.1–25. DOI: 10.1080/09588221.2020.1854311

El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems*, *38*(3), e12667, 1-15. https://doi.org/10.1111/exsy.12667.

Freedle, R., O. and Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. *TOEFL-Research Reports 44*, 5-56. Educational Testing Service.

Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system pigai. *Int. J. Emerg. Technol. Learn*. 16, 322–330. https://doi.org/10.3991/ijet.v16i11.19657

Gee, J. P. & Hayes, E. R. (2011). *Language and learning in the digital age*. Routledge.

Grabe, W. & Yamashita, J. (2022). *Reading in a second language: Moving from theory to practice (second edition).* Cambridge University Press.

Green, A. (2021). *Exploring language assessment and testing: Language in action (Second edition).* Routledge.

Golonka, E.M., Bowles, A.R., Frank, V.M., Richardson, D.L. & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27* (1), 70-105.

Hamuddin, J., B., Julita, K., Rahman, F., & Derin, T. (2020). Artificial intelligence in EFL context: Rising students' speaking performance with Lyra Virtual Assistance. *International Journal of Advanced Science and Technology*, *29*(5), 6735-6741. http://sersc.org/journals/index.php/IJAST/article/view/17726

Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.

Hughes, A. (2003). *Testing for language teachers (2nd edition)*. Cambridge University Press.

Jiang, R. (2022) How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology* 13, 1-8. https://doi.org/10.3389/fpsyg.2022.1049401

Kim, N. -Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence*, *17*(8), 37-46. https://doi.org/10.14400/JDC.2019.17.8.037

Kirsch, I. (2001). *The international adult literacy survey: Understanding what was measured*. RR-01-25. Princeton, Educational Testing Service.

Lai, K.-w. K., and Chen, H.-j. H. (2021). A Comparative Study on the Effects of a VR and PC Visual Novel Game On Vocabulary Learning. *Computer Assisted Language Learning*. *36* (3), 312-345. https://doi.org/10.1080/0958821.2021.192826.

Law, A. K., K., So, J., Lui, Ch, T., Choi, Y., F.,Cheung, K.,H., Hung, K.,K. & Graham, C.,A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education 25* (208), 1-9, https://doi.org/10.1186/s12909-025-06796-6

MacMillan, F., M. (2016). Assessing reading. In D. Tsagari & J. Banerjee (Eds.). *Handbook of second language assessment*, 113-129, Walter de Gruyter Inc.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. & Macalister, J. (2021). *Teaching ESL/EFL reading and writing*. Routledge.

Pöysäri, S., Siltala, N., Latokartano, J. (2025). Evaluating AI-genrated multiple choice exam questions in robotics education, *INTED2025 Proceedings*, 3119-3127. https://doi.org/10.21125/inted.2025.0809

Russel, S., J. & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Pearson Education.

Sadeghi, K. (2021). *Assessing second language reading. Insights from cloze tests.* Springer. ISBN 978-3-030-84469-1.

Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelli-gence in English language teaching and learning (2015-2021): What are the current effects? *Journal of Information Technology Education: Research, 21,* 337-377. https://doi.org/10.28945/4999

Tai, T.-Y., Chen, H. H.-J., and Todd, G. (2022). The impact of a virtual reality app on adolescent ELF learners' vocabulary learning. *Computer Assisted Language Learning*, 35 (4), 892–917. 10.1080/09588221.2020.1752735

van den Broek P. & Kendeou P. (2022). Reading comprehension I: Discourse. In M. J. Snowling Ch. Hulme& K. Nation (eds). *The science of reading: Handbook (second edition),* 239-260. John Wiley & Sons Ltd.

Yong, Q. (2020). Application of artificial intelligence to higher vocational English teaching in the information environment. *Journal of Physics: Conference Series, 1533* (3), 032030, 1-5. https://doi.org/10.1088/1742-6596/1533/3/032030

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech, 40*(1), 80-97. https://doi.org/10.1088/1742-6596/1533/3/032030