Chapter 8

A Comparative Analysis of Multiple-Choice Questions: ChatGPT-Generated Items vs. Human-Developed Items

Jean Y. Chun and Natalia Barley

Defense Language Institute Foreign Language Center (DLIFLC), Monterey, CA, USA

This study explores the potential of incorporating ChatGPT (GPT-3.5) to improve test development efficiency. It evaluates the quality of ChatGPT-generated multiple-choice questions (MCQ) compared to those written by human test developers. Additionally, the study seeks to identify the general characteristics of ChatGPT-generated items. A total of 80 items, 40 from ChatGPT and 40 from human writers, were developed from 20 authentic Korean passages. The quality of the items was evaluated by three raters on a five-point Likert scale against a rubric of eight criteria. Both quantitative and qualitative methods have been employed, incorporating an analysis of rating scores and the raters' written comments. The results indicate an overall comparability between ChatGPT-generated items and those created by human writers. However, ChatGPT's ability was significantly limited when creating plausible distractors. These findings underscore the importance of human judgment, particularly in the creation of effective distractors, to fully leverage ChatGPT in the test development process.

Introduction

Test development is a highly labor-intensive and time-consuming process (Kalpakchi & Boye, 2023; Shin & Lee, 2023), which challenges the production of high-quality tests and potentially threatens their validity as well. Validity, as one of the most important properties of tests, lies at the core of test usefulness. Construct validity is defined as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure"

Cite as: Chun, J. Y., & Barley, N. (2024). A Comparative analysis of multiple-choice questions: ChatGPT-generated items vs. human-developed items. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 118-136). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.08

(Bachman & Palmer, 1996, p. 21). In other words, scores obtained from tests that do not closely represent the intended construct fail to provide accurate information about test-takers' abilities, thereby undermining the original purpose of the tests. However, in reality, the development of tests that provide valid information for test users can be confined by another important property of tests—practicality. Practicality is "the extent to which the demands of the particular test specifications can be met within the limits of existing resources" (Bachman & Palmer, 1996, p. 36). The limitation of available resources can place a burden on language programs' efforts to develop tests that closely represent their intended construct. Striking a balance between validity and practicality poses a great challenge to many language institutes and programs, calling for a new approach to test development.

Recent developments in Artificial Intelligence (AI), however, show significant potential in overcoming the current challenge in language assessment. Among various advances, the introduction of Generative Pre-Trained Transformers (GPT), which is built on a Large Language Model (LLM), has garnered growing interest, particularly in language assessment. ChatGPT, a highly advanced chatbot, can engage in human-like interaction with users and respond to their requests (Kohnke et al., 2023). When applied in the context of language tests, ChatGPT can instantly generate multiple test passages and items prompted by test developers, serving as an automated content and item generator. Its ability has the potential to greatly enhance the efficiency of the test development process and mitigate resource constraints.

The potential of ChatGPT has sparked interest, leading to a series of studies examining its capability to generate test items. Attali et al. (2022) used GPT-3 to develop their reading test passages and items. The distractors for their selected response tasks, such as identifying the main idea and title of the given passages, were generated by selecting possible answers from alternative passages that share similar topics with the original passage yet differ in content. At least five raters reviewed each test passage and item in terms of its test validity and fairness. Of the initial 789 passages, 58% were retained for the test. The retained items exhibited a wide range of item difficulty, decent item discrimination, and local item independence.

Distinct from the previous study that focused on GPT items alone, Kalpakchi and Boye (2023) compared the abilities of various models, including GPT models (GPT-3 and ChatGPT), two fine-tuned models (Swedish Bert and SweCTRL-Mini) and baseline models, to produce Swedish reading multiple-choice questions (MCQs). The MCQs were evaluated by the researchers against the four criteria: (a) questions are grammatically correct, (b) the stem can be answered by the passage, (c) all answer choices are related to the stem, and (d) the answer choices contain a single correct answer. It was found that the GPT models produced much more acceptable levels of questions compared to the other models. Despite the overall outperformance of GPT models, especially in the production of grammatically correct items and items that can be answered using the information given in the text, GPT items also exhibited two recurring issues: (a) generating a key that was noticeably different from the distractors, potentially providing critical cues for test-takers, and (b) producing overlapping answer choices that resulted in multiple correct answers. These issues were more prevalent among GPT items than those generated by the other models.

Shin and Lee (2023) compared the passages and items written by human test writers with those generated by ChatGPT. They selected five questions from the English reading section of the 2019 College Scholastic Ability Test (CSAT) in South Korea and then generated five additional passages and MCQs by using ChatGPT. The CSAT and ChatGPT-generated passages and items were rated by 50 pre-and in-service teachers based on several criteria, including the naturalness of writing flow and expressions in passages, the attractiveness of answer choices, and the level of test item completion (item relevance to the passage, clarity of items, the homogeneity of options in content). The participants rated questions on a five-point Likert scale and provided their written rationale for their rating scores. They reported that the rating scores of reading passages written by CSAT, which were developed by human developers, did not differ significantly from those generated by ChatGPT. In contrast, the ChatGPT items received significantly lower scores than CSAT items in two areas: attractiveness of answer choices and item completion level.

Although these studies shed light on the characteristics of ChatGPT items, the knowledge regarding ChatGPT's ability to produce test items is quite limited. Furthermore, previous studies did not evaluate the quality of ChatGPT items based on a comprehensive list of criteria, failing to present a clear picture of what ChatGPT can or cannot do in terms of item generation. To embrace the use of ChatGPT in test development, the critical task lies in identifying areas in which ChatGPT excels and can assist human test developers in increasing overall efficiency in test development (Anders, 2023; Dell'Acqua, et al., 2023). To better understand the quality of ChatGPT items, this study evaluates the items with a comprehensive rubric and compares their quality to the quality of human writers' items. The findings of the study will help identify the areas where test developers can benefit from ChatGPT and those that require the judgment of human developers. The current study addresses two research questions as follows:

1. Does the quality of MCQ items differ between human item writers and ChatGPT?
2. What are the characteristics of ChatGPT-generated items?

Methods

This study was conducted at an intensive language program for military personnel. As one of the graduation requirements in this program, students must take the Defense Language Proficiency Test (DLPT), which consists of multiple-choice listening and reading comprehension questions. To help students prepare for this requirement, teachers are highly interested in developing practice tests used in the final semester. However, the development of high-quality MCQs is time-consuming, thus posing great challenges to teachers who have heavy teaching loads. There is a need to facilitate their test development process without compromising the quality of test items, which motivated the current study to find a way to support teachers by using ChatGPT.

Materials

The materials for the study were developed with existing Korean listening practice test items. A total of 40 practice test items with 20 passages (two items per passage) were selected from a pool of practice test items developed by a department in the Korean program. Teachers who constructed the practice items received one-hour training on writing multiple-choice questions from an experienced teacher prior to item development. Their items underwent an iterative process of peer review. Finalized items were edited by a native English speaker. The passages used in these items were authentic materials, selected from a Korean TV show called *Sebashi* (세바시). This show is equivalent to *TedTalk*, in which presenters share their thoughts and information on political and social issues. The selected passages were all rated at level 2/2+ on the Interagency Language Roundtable (ILR) scale; each passage contained two MCQs.

For comparison, the researchers used the same passages and generated two MCQs per passage by using ChatGPT (GPT-3.5). As the Korean program emphasizes the use of authentic instead of scripted materials for instructional and test preparation purposes, the passages were not generated by ChatGPT. The use of GPT-3.5 instead of 4.0 was due to the limitation in teachers' access to 4.0, which is a paid version. All teachers have free access to GPT-3.5, so the findings of the current study can benefit all faculty who use it. When generating items, ChatGPT received two pieces of information: (a) guidelines on writing multiple-choice questions—which were adapted from Haladyna's (2004) "general item-writing guidelines" (pp. 99 – 100) (Appendix A), and (b) revised Interagency Language Roundtable (ILR) skill level descriptors (Interagency Language Roundtable, n.d.). After entering the information, the following prompt was given to ChatGPT.

> I will give you an ILR level [...] passage. Create up to two ILR Level [...] multiple-choice questions in English based on the core information of the passage. Do not create items based on peripheral information. Make sure to follow the rules for creating multiple-choice questions that I gave you.

Since passages were rated either Level 2 or 2+, these levels were inserted within the brackets depending on the level of passages.

Procedures

The authors of the current study along with a faculty member from the Defense Language Institute participated in rating the test items. They were familiar with the ILR scale and test item writing. One of the authors and the faculty member conducted a blind review and the initial rating of the items, where the MCQs were presented in a random order without any indication of how the items were generated. The other author prepared the randomized items for the initial review. The raters evaluated the test items by using the eight-criteria rubric developed based on the MCQ development guidelines (Haladyna, 2004). The rubric is presented in Table 1.

Table 1

Test Item Rating Rubric

	Criteria
1	The stem and answer choices do not contain misleading or ambiguous language. They are written clearly.
2	The stem and answer choices target core information (not peripheral) such as the main idea or key details of the passage.
3	There is a single correct answer.
4	There is a clear correct answer.
5	The correct answer is indicated accurately.
6	The answer choices are distinct from one another.
7	All stems and answer choices are plausible. It is hard to identify the correct answer without reading.
8	All items (stems and answer choices) are distinct from one another, not having overlapping ideas between them.

Note. Criterion No. 3 is concerned with the absence of multiple correct answers; Criterion No. 4 is about whether raters can clearly identify the correct answer within the passage.

The test items were evaluated on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). When the raters gave a score of 3 or lower, they provided their written rationale. After completing the rating, the absolute agreement rate was calculated. The rate, at 79.33%, was considered acceptable according to Stemler (2004). The items with a significant discrepancy in scores on any given criterion were discussed to reach an agreement. After discussion, all items received a single rating score.

Data Analysis

This study's data comprises the final rating scores assigned to rubric items, along with the accompanying written rationales. Both quantitative and qualitative analysis methods were conducted to analyze the data. To answer the first research question regarding the comparison between ChatGPT items and human writers' items in rating scores, descriptive statistics and the Wilcoxon signed-rank test were first performed on the entire data set, including the rating scores of ChatGPT and human test items. Given that the data comprises ranked responses, which is ordinal data, and considering that the rating scores were provided by the same raters for both human writers' and ChatGPT's items, the Wilcoxon signed-rank test (a non-parametric analysis using the matched sample) was selected for analysis (Schulyer, 2012). After reviewing the descriptive statistics, the Wilcoxon signed-rank test was performed again for the rubric criteria that showed a noticeable difference in rating scores between ChatGPT and human developers. To address the second research question, which pertains to the general characteristics of ChatGPT items, both descriptive statistics on the rating scores of ChatGPT items and the written rationales provided by the raters were employed for analysis. The raters' comments provided for the ChatGPT-generated items were separated from the comments on human developers' items. A content analysis of the comments on ChatGPT items was then performed to identify the types of errors in ChatGPT-produced items and the frequency with which they occurred.

Results

The first research question examines whether the quality of items differs significantly between ChatGPT and human writers. The descriptive statistics showed that the quality of test items was comparable overall (Table 2).

Table 2

Descriptive Statistics of Test Item Quality

		ChatGPT		Human	
	Criteria	Mean	*SD*	Mean	*SD*
1	Clear language	4.80	0.76	4.51	0.85
2	Core Info	4.90	0.63	4.85	0.70
3	A single answer	4.33	1.49	4.15	1.55
4	A clear answer	4.23	1.58	4.28	1.54
5	Accurate indication of an answer	4.50	1.34	4.85	0.66
6	Distinct options	4.70	0.85	4.80	0.61
7	Plausibility	2.93	2.00	3.83	1.74
8	Item independence	4.35	1.39	4.60	1.05
	Average	4.34	1.44	4.48	1.21

The average scores for ChatGPT and human writers were 4.34 and 4.48, respectively. The results of the Wilcoxon signed-rank test showed that there was no significant difference between them (p = .12). Among the rubric criteria, criterion No. 7 showed a noticeable difference in rating scores between ChatGPT and human items. The average score for ChatGPT items was 2.93, while human-writer items received 3.83 on average. The Wilcoxon signed-rank test was performed again to examine the significant difference between them. It was found that ChatGPT items received significantly lower rating scores in this criterion than human writers' items (p <.05), which suggests that the distractors generated by ChatGPT tend to be perceived as less plausible than those written by human developers.

To answer the second research question, a content analysis of the written justifications for the ratings was conducted. Since justifications were provided only for the rubric criteria rated 3 or below, descriptive statistics were referenced to determine the areas of strength of the

ChatGPT-generated items (Table 2). Several patterns were identified in the items developed by ChatGPT.

ChatGPT consistently generated items that target core information from the passages, rather than peripheral information (*M*=4.9, *SD*=0.63) and contain virtually no misleading or ambiguous language (*M*=4.8, *SD*=0.76). Most items also contained answer choices that are distinct from one another, i.e., do not contain overlapping ideas (*M*=4.7, *SD*=0.85) and there is a single, clearly and accurately identified key (*M*=4.2-4.5, *SD*=1.3-1.6). The items provided below were produced by ChatGPT based on a passage about immigration (Appendix B). (The keys are bolded.)

1. What is the main concern expressed in the passage regarding the number of university students in South Korea?
   A) The shortage of foreign exchange students.
   B) The potential impact on the domestic economy.
   C) The lack of available jobs for graduates.
   D) The need for more universities.

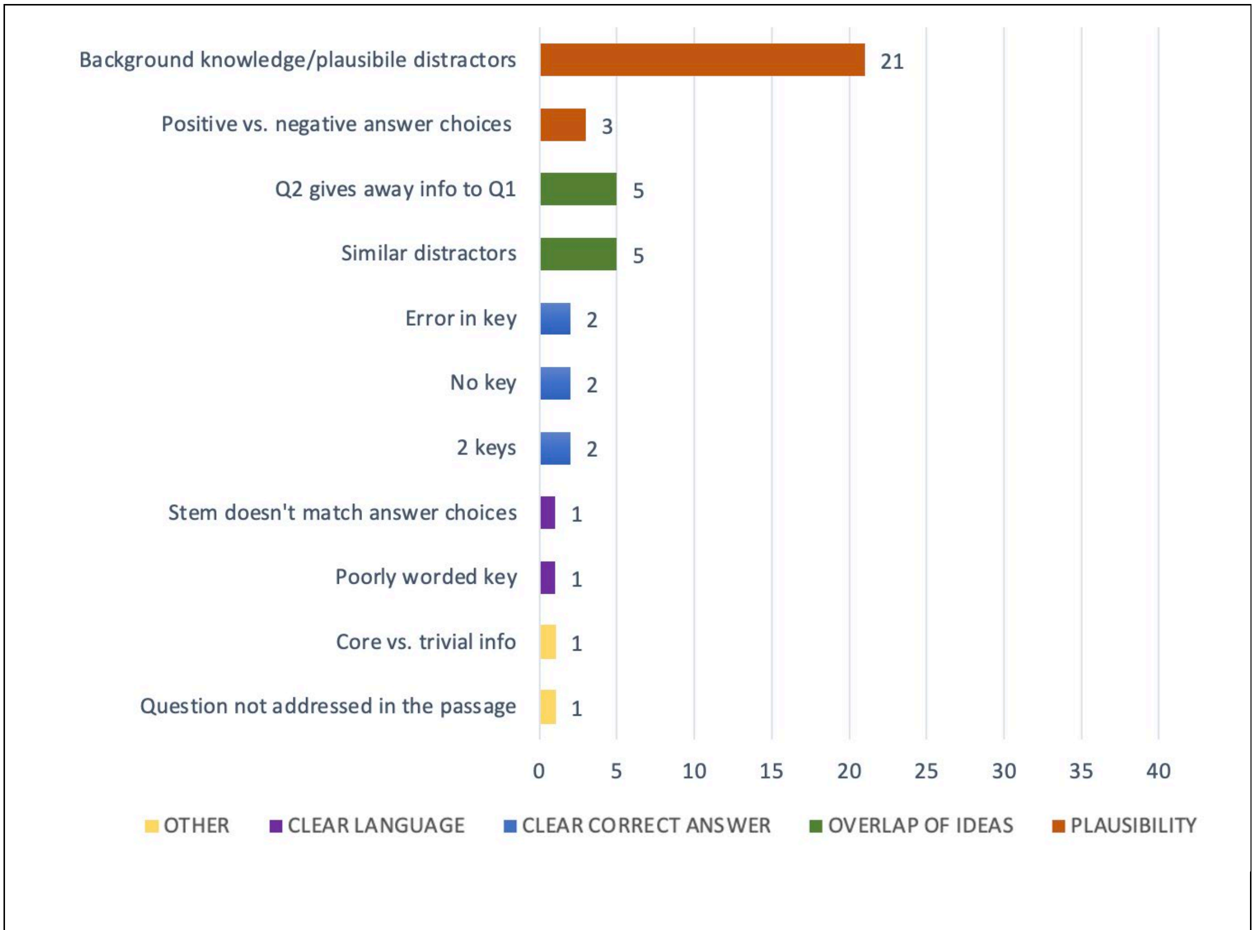
2. Why did the speaker mention the scenario involving sending all foreign nationals out of the country?
   A) To criticize a proposal made by someone in the past.
   B) To suggest a potential solution to the job shortage.
   C) To highlight the importance of international tourism.
   D) To emphasize the need for stricter immigration policies.

These items were reviewed against the eight abovementioned criteria with no identified errors. They target essential content in the passage and do not focus on peripheral information. The speaker in the passage discussed the declining number of university students in Korea, emphasizing an anticipated drastic decrease in the student population in the future, which could potentially have a devastating impact on the domestic economy. This premise, which is critical to understanding his main point of promoting immigration, is well captured in the first item. The second item, in turn, asks about the speaker's intention behind addressing the scenario to assess test takers' understanding of the speaker's concern against those opposing immigration, distinctly separate from the first item. Besides targeting two independent core aspects of the passage, the stems and distractors also contain neither grammatical errors nor ambiguous language. A single, clear key is indicated by ChatGPT as well. These items received the highest rating, a score of 5, on each criterion from raters, highlighting the positive characteristics of ChatGPT items.

The content analysis of the raters' justifications revealed that the most frequently-encountered error is related to the plausibility of distractors (Figure 1). More specifically, the raters could often identify the key based on the background knowledge of the subject in question and the general world knowledge without reference to the passage. The comments regarding background knowledge were observed in almost half of ChatGPT items (21 out of 40, 52.5%).

Figure 1

Content Analysis Summary



The example below received the following comment from one of the raters: "A & C are not plausible. The correct answer can be guessed using background knowledge." This comment suggests that test takers can likely identify the correct answer using their general knowledge of adolescent behavior and familiarity with bullying as a social phenomenon.

What was the speaker's primary motivation for wanting to help children who experience bullying?

A) To gain recognition and popularity among peers.
B) To overcome personal difficulties and traumas.
C) To become a leader in their school community.
D) To explore the reasons behind bullying behavior.

Similar to the example above, the following example also elicited comments from raters indicating that their background knowledge helped them narrow down their choices. One of the raters stated that she was easily able to eliminate some distractors with her background knowledge ("A & B can be eliminated using background knowledge."). Another rater's comments also suggest that her background knowledge played a critical role in detecting unattractive distractors ("Why would influential individuals participate in rumors first? It can easily destroy their career. Also, the public would not have the expertise to spread general information").

> What structural pattern distinguishes the spread of rumors from the dissemination of general information?
>
> A) Rumors typically start with influential individuals and spread to the general public.
> B) General information originates from the general public and spreads to influential individuals.
> C) **Rumors exhibit a decentralized structure with weaker connections.**
> D) General information forms a centralized structure with strong connections.

Another identified issue that impedes the plausibility of distractors is related to how some distractors are structured by ChatGPT. In several instances (*n*=3), the key further stood out as the answer choices were structured to highlight the "positive [distractors] vs. negative [the key] aspect of the problem," as indicated by one of the raters. The examples below demonstrate that point. Three distractors are all positive, whereas only the answer key is negative, which provides a clear hint for test-takers to guess the correct answer.

> 1. According to the passage, what is a common outcome for students who engaged in violence during their school days?
>
> A) They often excel academically and professionally. [Positive]
> B) They tend to become successful in their careers. [Positive]
> **C) They may face difficulties in various aspects of life, including academics and employment.** [Negative]
> D) They generally have a strong support system to help them overcome challenges. [Positive]
>
> 2. According to the passage, why do Korean students study extensively despite having the least interesting education globally?
>
> A) Korean students are naturally more motivated to study than students from other countries. [Positive]
> B) Korean students are driven by their passion for academic excellence. [Positive]
> **C) Korean students study out of fear of future challenges and consequences.** [Negative]
> D) Korean students have access to superior educational resources. [Positive]

The errors observed much less frequently were overlapping ideas between item pairs or distractors (e.g., similar distractors, information in one item giving clues to the other item related to the same passage), unclear correct answers (e.g., two or zero keys in an item, errors in keys), a poorly worded key, the stem not matching the answer choices, a question targeting trivial (vs. core) information, and a question targeting information not addressed in the passage. While many items contained only one error, only 30% of the ChatGPT-generated items had no identified issues.

Below are examples of the less-frequent errors. The following two questions, which are based on a passage about education in Korea, present several challenges. For example, the stem of the second question gives a clear clue for the first question, indicating that answer choice A in the first question is the likely correct answer. The answer choices C and D in the second question communicate the same information, subsequently reducing the number of distractors.

1. What is the main concern the speaker expresses about education in their society?
   A) Students lack a sense of individuality and diversity in their dreams.
   B) Students are not receiving proper counseling from their teachers.
   C) Students are too focused on achieving similar goals without variety.
   D) Students often fail to understand the importance of education.

2. According to the passage, what positive outcomes have arisen from embracing diverse individuals who were once considered “strange” or “different”?
   A) They have created imaginative and innovative concepts.
   B) They have established a unified cultural identity.
   C) They have adhered to traditional norms and values.
   D) They have conformed to societal expectations and standards.

In the following example, the answer choice A was marked by ChatGPT as the key; however, the passage provides the speaker’s account of an accident she was involved in, rather than witnessed (see Appendix C). The remaining answer choices are also incorrect.

What happened to the speaker during their commute home one summer day?
   A) She witnessed a car crash.
   B) She encountered heavy traffic.
   C) She got lost on the way.
   D) She had a flat tire.

Another issue related to answer keys is what the raters labeled as “errors” in the otherwise correct answers. Both “errors” involved a reversal of terms. In the following example, there is an error in the answer key C where the word “spending” is used instead of the correct “saving.” (See Appendix D for the passage.)

What is the shift in the speaker’s perspective about saving money and spending habits?
   A) She now sees saving money as the main goal.

B) She believes that spending is essential for happiness.
C) She considers **spending** as a way to earn more.
D) She views spending as an unnecessary burden.

In the example below, the definitions were reversed in the key. The correct answer would be: "Academy-led learning relies on external judgment, while self-directed learning relies on self-assessment for learning needs." The issues of distractor plausibility and overlapping ideas are also evident here.

What is the key difference between academy-led learning and self-directed learning as described in the passage?

A) Academy-led learning relies on self-assessment for learning needs, while self-directed learning relies on external judgment.
B) Academy-led learning involves asking oneself questions, while self-directed learning involves seeking guidance from others. s
C) Academy-led learning emphasizes formal education, while self-directed learning focuses on informal learning.
D) Academy-led learning involves academic subjects, while self-directed learning involves vocational training.

Discussion

This study examines the differences in the quality of MCQ items produced by human item writers and ChatGPT and analyzes the characteristics of ChatGPT-generated items. Many of the findings are consistent with the results of the previous studies. Similar to Kalpakchi and Boye's (2023) findings, most items generated by ChatGPT were written clearly and were grammatically correct. Both studies also found that most of the stems targeted information provided in the text. In investigating clearly marked, single correct answers, Kalpakchi and Boye (2023) also found a few items with no key and some with overlapping options, i.e., more than one possible correct answer.

The analysis of GPT-3 items by Attali et al. (2022) identified very few analyzed item pairs (i.e., items derived from the same passage) with local item dependence indicating that most items on the test are not related to one another. Although this study did not pilot the test items or conduct psychometric analysis, the raters reported similar observations with only five item pairs giving away clues between questions.

Similar to the present study, Shin and Lee (2023) identified the quality of distractors as the most commonly reported issue when compared with human-developed items, although the study did not focus on the analysis of what specifically made the distractors "unattractive" (p. 34). Interestingly, Kalpakchi and Boye (2023) found few GPT-produced items that are answerable without reading the passage, in contrast to 52.5% of such items in this study. Instead, they observed many keys that "stick out", i.e., are noticeably different from the others, providing cues for test-takers compared to only three such items ("positive" vs. "negative" answer options) in the present study.

ChatGPT demonstrates a remarkable potential in assessing reading and listening comprehension by generating MCQ item stems that target core information and by correctly identifying the key. However, creating plausible distractors has proven to be a much more challenging task that requires a nuanced understanding of the test takers' world and background knowledge. For that, human judgment is essential. Studies investigating the use of GPT in other domains also found that in tasks that call for human judgment and interpretation, AI tools demonstrated poorer performance (e.g., Dell'Acqua et al., 2023).

Tasks that fall within the "technological frontier" (Dell'Acqua et al., 2023) of ChatGPT are those that follow structured and rule-based processes, enabling the model to operate within clearly defined instructions. However, when faced with the inherently subjective task of generating distractors that demand context-specific relevance and appropriateness, such as considering the background knowledge of test-takers in developing plausible distractors, the model faces difficulties. The necessity for a nuanced understanding, context assessment, and discernment in crafting realistic and challenging distractors highlights the constraints within the current technological boundaries of ChatGPT.

It is important to note that the expectation, at least with the currently available generative transformers, is not to fully automate item generation. Instead, the goal is to augment and enhance human performance with AI. With the "human in the loop" (Wang, 2019) of the design process, we can capitalize on the efficiency of intelligent automation while engaging in conversation with it, providing intermediate feedback, and curating an output in line with the desired intention. What requires further investigation is whether generative transformers can create plausible distractors with appropriate human feedback or whether this task is best accomplished without AI augmentation.

Conclusion

This study aimed to investigate the suitability and capability of ChatGPT for multiple-choice question (MCQ) generation, addressing two key research questions. Overall, the findings revealed that the sets of MCQ items generated by human item writers and ChatGPT were largely comparable across various criteria, such as language clarity, core information targeting, and the identification of a clear, single correct answer. Both methods were effective in developing distinct distractors and ensuring item independence. However, a notable distinction emerged in the quality of distractors, with more than half of the ChatGPT-generated items featuring implausible distractors. These distractors, answerable through world or background knowledge alone, raised concerns about the tool's ability to create challenging and context-dependent options.

While ChatGPT demonstrates promising potential in crafting MCQ stems and keys, the study underscores the indispensable role of human judgment in generating plausible distractors.

Tasks that fall within ChatGPT's current capabilities align with structured and rule-based processes, allowing it to navigate within the confines of explicit instructions. Yet, when confronted with the inherently subjective nature of crafting distractors that require context-specific relevance and appropriateness, the model encounters challenges. The need for nuanced understanding, context evaluation, and discernment in the creation of realistic and challenging distractors underscores the limitations of the current capabilities of ChatGPT.

As the field of generative AI continues to evolve, future research should concentrate on enhancing the plausibility of distractors through meaningful human-AI interactions. Exploring a variety of prompts for ChatGPT, prompting the model to engage in iterative reviews with targeted feedback, and experimenting with diverse strategies for refining distractors could be valuable avenues for further investigation. By addressing these aspects, researchers can contribute to the ongoing development of AI-powered tools in educational settings, ensuring a harmonious collaboration between human expertise and artificial intelligence capabilities in the realm of MCQ generation.

References

Anders, A. (2023, October 1–21). *Generative AI: Experiments in Writing and Learning* [Plenary Talk]. Technology for Second Language Learning, 2023, Ames, IA. Retrieved November 7, 2023, from https://youtu.be/kDmLdIJ_xYI

Attali, Y., Runge, A., LaFlair, G.T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 1–13. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013). https://www.iab.cl/wp-content/uploads/2023/11/SSRN-id4573321.pdf

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge.

Interagency Language Roundtable. (n.d.) *Revised Interagency Language Roundtable Language Skill Level Descriptions*. Retrieved September 23, 2023, from https://govtilr.org/Skills/Listening.htm

Kalpakchi, D., & Boye, J. (2023). Generation and evaluation of multiple-choice reading comprehension questions for Swedish. Retrieved October 25, 2023, from https://www.diva-portal.org/smash/get/diva2:1771278/FULLTEXT02

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2). https://doi.org/10.1177/00336882231162868

Schulyer, W. H. (2012). *Reading statistics and research* (6th ed.). Pearson.

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

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, *9*(1), 4. https://doi.org/10.7275/96jp-xz07

Wang, G. (2019, October 20). *Humans in the loop: The design of interactive AI systems.* Stanford University. Retrieved November 10, 2023, from https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

Appendix A

Multiple-Choice Question Writing Guidelines for ChatGPT

1. The stem and answer choices must not contain misleading, or ambiguous language. They must be written clearly.
2. The stem and answer choices must target core information (not peripheral). Core information refers to the essential or fundamental details, facts, or concepts that are crucial to understanding a particular passage. The examples of core information are the main idea or key details.
3. There must be a single correct answer.
4. There must be a clear correct answer. A test-taker should be able to identify a correct answer clearly.
5. The answer choices must be distinct from one another.
6. All options must be plausible. It should be hard to identify the correct answer without listening.
7. All items (stems and answer choices) must be distinct from one another, not having overlapping ideas between them.
8. Do not use negatives in your stem.
9. Indicate a correct answer accurately.
10. Maintain similar answer choice lengths.
11. Do not use absolute language such as “every”, “always”, “solely” or “only”.

Appendix B

English Translation of Korean Passage on Immigration

Shall we briefly consider international exchange students as well? As of 2021, there are 333 universities in our country, with an admission capacity of 480,000 students. As I mentioned earlier, the number of newborns is 270,000. Even if all of them enter universities after 18 years, the admission capacity would still be 270,000. In other words, there would be a remaining 210,000 students. If we calculate the admission capacity of each university as 1,400, then after 18 years, it means that 140 universities would not be able to have new students.

The closure of a university not only affects the university itself but also completely devastates its surrounding areas. To maintain the existing universities, we would need 210,000 international exchange students annually. Naturally, they would have a significant impact on the domestic economy.

However, there is currently a highly controversial issue. It is the claim that foreigners are taking away our jobs. Let me share a story related to employment. A few years ago, there was someone arguing that immigration policies were flawed, especially because they believed that jobs were being lost because of immigrants. His conclusion was to send back all foreign workers in the country, claiming that they were taking away our jobs. I asked him, "Based on your question, are you suggesting that all countries in the world should send back the foreigners in their respective countries?" He answered yes. So, I said to him, "In that case, if all 7.5 million overseas Koreans return to our country and if we also send back all the travelers and exchange students, which would total 2 million residents, how would our job market improve?" He hung up the phone, and I never heard from him again. In reality, about 10% of small-scale manufacturing industries claim to have a shortage of workers. Regarding the hiring of foreign workers, 81.3% responded that it was to alleviate the labor shortage.

Appendix C

English Translation of Korean Passage About the Automobile Accident

I was in my fourth year of university when I was involved in a car accident on my way home from school during the summer. My older brother and I were driving along our usual route, planning to go to our parents' house once the traffic light turned green. However, a driver under the influence of alcohol approached from behind and, instead of applying the brakes, accelerated and collided with the back of our car. The impact was so strong that our small car crossed the center line and collided with seven other vehicles, eventually catching fire. As a result, the flames engulfed my body, and I lost consciousness. My brother also suffered severe burns on his arm as he pulled me out. That night, the ordinary journey home for my brother and me turned into something far from ordinary, to the extent that it made the news. After that incident, I experienced several near-death situations, and it took seven months before I could finally return home.

"I **had** [당했습니다] an accident (this verb choice [very common in Korean] indicates that the speaker received this action; someone caused this action to the speaker, and the speaker is the byproduct of this accident.)." That's the correct term to use. However, at some point, it became uncomfortable for me to say, "I had an accident." Every time I used that phrase, it felt like I was labeling myself as a "victim of a drunk driving accident." Regardless of the wording I choose to describe the incident, the fact remains that an unexpected accident occurred on that ordinary night, and as a result, I suffered tremendous harm. But at least I didn't want to be defined solely as a victim, as I had gone through the time after that night in order to survive.

Appendix D

English Translation of Korean Passage on Money

Before I started investing [personal finance management], when I was a YOLO*, I thought this saving was very pathetic. But as I started investing [personal finance management], my thoughts changed. ‘It’s not about saving, it’s about earning.’ Now, my standards for consumption have become very high, and my ability to control my spending is at its peak. I don't feel the desire to buy random things anymore. In the past, I would have thought, ‘Oh, it’s pretty, I should buy it,’ but now I just pass by, thinking, ‘It’s not something I need.’

So, I started thinking, ‘It’s not about saving 10,000 won, it’s about earning 10,000 won.’ Some people say to me, “You have an amazing mental strength”; “You have amazing patience” [a mocking tone] or “Even if you gather every speck of dust, it is a mere collection of dust.” But I can confidently say, ‘Only those who gather specks of dust can create a huge mountain.’ I mentioned that my ability to control spending has improved. The most helpful thing in this regard was raising the standard of essentials. Before buying something, I ask myself three questions to raise the standard of essentials.

Actually, yesterday when I went to the mart, I saw a scalp massager that provides a scalp massage while shampooing. Naturally, I thought, ‘Oh, I want to buy it,’ but I immediately stopped myself, thinking, ‘I should raise the standard of essentials.’ With this thought in mind, I asked the first question, ‘Is it truly a necessary item?’ In fact, just asking this question alone helps filter out most unnecessary things. Thoughts like, ‘Wouldn’t it be nice?’ or ‘Isn’t it something I need?' come up. Instead of buying based on thoughts like, ‘If the scalp feels cool, wouldn’t it improve scalp health and hair texture?’, I move on to the second question.

*Note: “욜로족(Yolo-jok)” refers to individuals who live in the moment and spend extravagantly.

About the Authors

Jean Young Chun

Jean Young Chun obtained her doctoral degree in Second Language Studies from Indiana University, Bloomington in 2022. She currently serves as an assessment specialist in the Curriculum Support (CS) Division at the Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, CA. Her previous career focused on language teaching in secondary and post-secondary institutions. Her research interests center on L2 assessment and pragmatics. Her dissertation investigated methods to assess L2 learners' comprehension of implied meanings by evaluating their immediate oral responses.

Natalia Barley

Natalia Barley holds a doctoral degree in Learning Design and Leadership and currently serves as the Director of the Curriculum Support Division within the Educational Technology and Development Directorate at the Defense Language Institute Foreign Language Center (DLIFLC). Her prior assignments include department chair in the basic, intermediate, and advanced language programs, a curriculum specialist, a faculty development specialist, an OPI tester, and a language instructor. Her research interests include technology-mediated language learning, with a focus on negotiation of meaning in synchronous online environments, as well as technology integration in language teaching and learning.