Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

October 2023, Volume 27, Issue 3
pp. 27–40

LANGUAGE TEACHER EDUCATION AND TECHNOLOGY FORUM

Can ChatGPT make reading comprehension testing items on par with human experts?

Dongkwang Shin, *Gwangju National University of Education*
Jang Ho Lee, *Chung-Ang University*

Abstract

Given the recent increased interest in ChatGPT in the L2 teaching and learning community, the present study sought to examine ChatGPT's potential as a resource for generating L2 assessment materials on par with those created by human experts. To this end, we extracted five reading passages and testing items in the format of multiple-choice questions from the English section of the College Scholastic Ability Test (CSAT) in South Korea. Additionally, we used ChatGPT to generate another set of readings and testing items in the same format. Next, we developed a survey made up of Likert-scale questions and open-ended response questions that asked about participants' perceptions of the diverse aspects of the target readings and testing elements. The study's participants were comprised of 50 pre- and in-service teachers, and they were not informed of the target materials' source or the study's purpose. The survey's results revealed that the CSAT and ChatGPT-developed readings were perceived as similar in terms of naturalness of the target passages' flow and expressions. However, the former was judged as having included more attractive multiple-choice options, as well as having a higher completion level regarding testing items. Based on such outcomes, we then present implications for L2 teaching and future research.

Keywords: *Artificial Intelligence, Automated Item Generation, ChatGPT, Content Generation*

Language(s) Learned in This Study: *English*

APA Citation: 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

Introduction

Chat Generative Pre-trained Transformer (ChatGPT, henceforth), which is a Large Language Model (LLM)-based chatbot, has received significant attention in a wide range of domains in recent months. Since its release on November 30, 2022, ChatGPT reached one million users in five days and two million in two weeks, thus expanding quicker than the record-breaking Facebook (now Meta), which took ten months to reach 100 million users (Kim, 2023). Although no large-scale, experimental research has yet reported on ChatGPT use in the language learning and teaching field, some studies (e.g., Ahn, 2023; Kohnke et al., 2023; Kwon & Lee, 2023; Shin, 2023) have already started revealing its full potential in the language teaching and learning domain.

Among various ways that ChatGPT can be employed in language teaching and learning, we focus on its capability to generate original texts and language testing items in this article. Although the concepts of computer-based text generation and automated item generation were introduced more than five decades ago (e.g., Bormuth, 1969; Klein et al., 1973), it is only recently, with the development of LLM (i.e., ChatGPT), that such technology has reached a level where it could be a useful supplement to language teaching and learning (Shin, 2023). Given that generating original passages and designing testing items is labor-intensive for L2 teachers, we explore ChatGPT's potential to this end. More specifically, we address the question of whether ChatGPT could make L2 readings and tests on par with those generated by human experts through a blind test, in which both pre- and in-service teachers evaluated different

aspects of the College Scholastic Ability Test (CSAT) (i.e., the Korean SAT) and ChatGPT-generated content. Based on our findings, we also present the implications for future research and L2 teaching in the context of using ChatGPT.

Background

In this section, we review three relevant branches of research that relate to our topic: research on computer-based content generation, research on automated item generation, and recent education studies on ChatGPT.

Research on Computer-based Content Generation

Developing computer-based content generation systems dates to the 1970s with the introduction of the Novel Writer System (NWS) by Klein et al. (1973). The NWS was employed to draft a short story when a user entered basic information about the story's historical background and characterization. During the same decade, Meehan (1977) produced Tale-Spin, an interactive storytelling program that allowed users to enter a sequence of events and receive a story in return that incorporated those elements. Despite their early advancements, however, these early systems were limited in that they relied on rule-based or case-based methods. That is, they required a significant amount of human effort to categorize story components, such as characters, scenes, and plot.

More recently, automated content generators based on artificial neural networks and machine learning have been developed. These systems can automatically produce stories using basic information like setting, characters, and plot. The LLM that has become the backbone of such neural network-based content creators is GPT-3 (Generative Pre-trained Transformer 3). GPT-3 is an artificial intelligence language prediction model constructed by OpenAI and has gained extraordinary attention in myriad domains (Ghumra, 2022). Trained on a dataset of over 200 million English words, GPT-3 can generate sophisticated English sentences and perform diverse language-related tasks, all based on users' prompts. Such an advanced technology has led to the burgeoning of various AI-based text generation tools like CopyAI, Hyperwrite, and INK. In one recent study (Lee et al., 2023), a reading activity created by an AI-based content generator has been found to enhance young EFL learners' English reading enjoyment and interest in reading English books, thereby underscoring the value of adopting such technology in L2 classrooms.

Research on Automated Item Generation

Automated item generation refers to the process of automatically drafting testing items using measurement information (Bormuth, 1969). Moreover, it could be considered that such technology could be extremely useful in terms of time and cost in the language testing and learning domain, especially when building large-scale testing banks.

To date, automated item generation methods have been divided into largely two approaches: ontology-based and rule-based. The former is a model that represents an object or concept in a form that can be understood by both humans and computers, and it focuses on the object's properties or relationships (Al-Yahya, 2014). Utilizing an ontology-based approach, Al-Yahya (2014) automatically generated multiple-choice (MCQ), true/false (T/F) and fill-in-the-blank (FB) questions. The process was restricted to drafting toss-up and wrong answers. Meanwhile, Liu et al. (2012) implemented the rule-based approach. They developed a program for scientific writing called G-Ask, which can produce prompts for scientific writing.

More recently, deep-learning-based automated item generation approaches have become mainstream. One type of Recurrent Neural Network (RNN) deep learning technique, the Long Short-Term Memory (LSTM) network, was used by von Davier (2018) to produce items measuring non-cognitive qualities (e.g., emotions, attitudes). Kumar et al. (2018) also employed an LSTM model to analyze inputted sentences and create question–answer pairs for them. ChatGPT, which was used for automated item

generation in this study, also uses a deep learning approach. However, it could be considered superior to earlier technologies because it employs an LLM-based transformer method (Shin, 2023).

Education Research on ChatGPT

In this section, we provide a selective review of recent studies that explore the following issues: the use of ChatGPT in L2 reading and its potential to replace human teachers. First, Kwon and Lee (2023) analyzed ChatGPT's accuracy in answering English reading comprehension questions on the CSAT and the TOEFL iBT. The results showed that ChatGPT, based on GPT-3.5, was indeed capable of answering approximately 69% of the questions correctly. In terms of question type, ChatGPT correctly answered about 75% of those pertaining to factual and inferential comprehension and around 87% of the fill-in-the-blank and summarizing ones. However, its accuracy rate on vocabulary and grammar questions was much lower. Nonetheless, ChatGPT PLUS, which is based on the latest GPT-4, achieved 93% mastery, including on vocabulary and grammar questions.

With a similar goal, Ahn (2023) evaluated ChatGPT's efficacy on CSAT English reading comprehension testing items. In the experiment, ChatGPT provided correct answers 74% of the time. The study therefore suggests that ChatGPT's performance could be improved by enhancing training methods, incorporating diverse and balanced datasets, and applying human-AI collaboration. The research also identified testing item types that ChatGPT could advance, including identifying the most appropriate order of events in a story, determining pronoun referents, and placing sentences in correct order.

While prior literature focused on measuring ChatGPT's capability to solve reading comprehension problems as the test taker, Shin (2023) investigated ChatGPT's potential in terms of developing reading comprehension items as the test designer. The outcomes showed that some question types require specialized prompts for them to be designed properly; therefore, the different kinds of questions included identifying the contextual meaning of underlined expressions, sequencing parts of a passage, and identifying the mood of the text. Based on the measurements, the study provided optimized prompts for different types of reading comprehension questions, as well as for various question development tips when using ChatGPT.

Despite growing recognition of ChatGPT's capabilities in L2 research, there appears to be notable apprehension among teachers regarding their roles being potentially replaced by AI. For example, in a study conducted by Chan and Tsi (2023), a survey was administered to 384 undergraduate and graduate students, as well as to 144 teachers representing various disciplines. The goal was to gain insight into the potential of AI, including ChatGPT, to replace teachers. The results showed that neither group strongly agreed that AI could replace teachers in the future. In another study, Tao et al. (2019) reported on the results of a survey conducted among 140 teachers and found that about 30% expressed skepticism about AI replacing them. However, the remaining believed otherwise.

In the current state of mixed skepticism and optimism about AI, including ChatGPT, it is important to note that the L2 teaching field needs classroom-oriented research that could demonstrate the technology's usefulness to L2 teachers and practitioners. Based on the results of such research, they can then judge its potential for themselves. To address this issue, in this study, both in- and pre-service teachers in the Korean EFL context were asked to evaluate reading passages and questions generated by the CSAT and ChatGPT in a blind test, which aligned with the following two research questions:

Research Question 1. How do pre- and in-service English teachers perceive the CSAT and ChatGPT-developed reading passages in terms of the naturalness of the writing flow and the expressions?

Research Question 2. How do pre- and in-service English teachers perceive the CSAT and ChatGPT-developed reading comprehension testing items in terms of the attractiveness of multiple-choice options and the overall level of completion?

Method

This section describes the methodological approach of the blind test, during which participants were asked to evaluate different aspects of the CSAT and ChatGPT-developed reading passages and testing items.

Participants

The present study included two groups of participants. The first was comprised of 38 undergraduate students majoring in English Education (pre-service teachers, henceforth) at a private university in Seoul, South Korea. These participants had been preparing for the teacher's examination to qualify as in-service teachers, and they had already taken several courses related to English Education. In this group, about two-thirds (68%) had taken a course on L2 assessment and testing. The other group was comprised of 12 in-service English teachers and professors (in-service teachers, henceforth), with a mean career length of 10.6 years (*SD* = 5.2). Each group's self-judged rating of L2 reading assessment, as found in the survey's background section, was 3.08 and 3.45 for the pre- and in-service group, respectively, with 5 being the most proficient and 1 being the lowest.

The participants were not informed of the present study's aim (i.e., to evaluate ChatGPT-developed reading passages and testing items) but were told only that they had been asked to evaluate items on the CSAT's English test (reading section). Two participants (one in each group) had not completed the survey and were therefore removed from the analysis.

Description of the Technology regarding Text and Item Generation

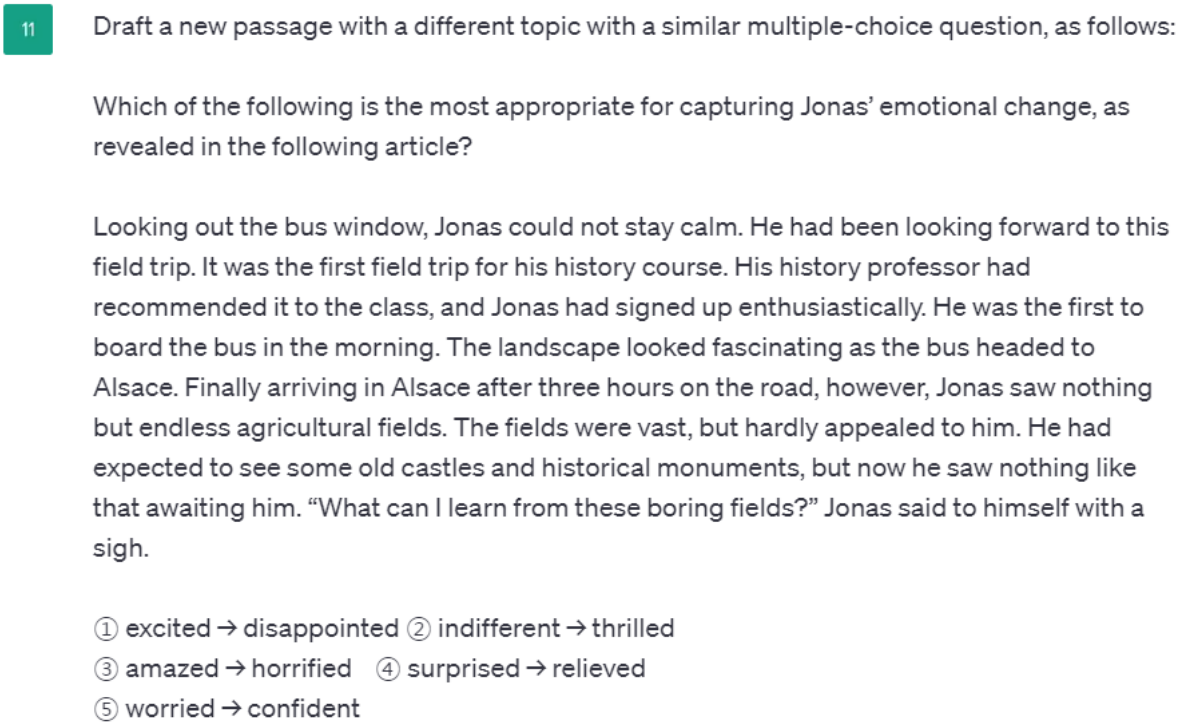
In this study, we used ChatGPT to generate items based on those of the CSAT English test in South Korea. Based on Shin's (2023) suggestion that using model items is more effective than relying solely on prompts to create test components, we selected five reading questions from the English section of the 2019 CSAT (Korea Institute for Curriculum and Evaluation, 2019) as model items. These five items were comprised of different question types, including (1) identifying changes in a character's mood, (2) pinpointing details in a passage, (3) inferring an appropriate phrase from blanks, (4) inserting a paragraph according to the text's flow, and (5) filling in appropriate words in the blanks provided in a given passage's summary. After selecting the five items of different question types from the CSAT, we then used ChatGPT to generate a comparison set of five reading passages and testing items. For example, as shown in Figure 1, the reading passage and testing item related to the question type 'identifying changes in the character's mood' from the CSAT were typed into ChatGPT along with the following prompt: *Draft a new passage with a different topic with a similar multiple-choice question, as follows.*

A similar procedure was followed for the other kinds of question, although some modifications were made to the prompts for each type. That is, if the produced testing item differed from its CSAT counterpart in terms of the target question type's structure (or format), the prompts were slightly revised. For instance, in the case of the 'pinpointing details in a passage' question type, the following prompt was entered: *Draft a new passage with a different topic and a multiple-choice question to confirm the agreement with the details in the passage*. Such procedures were continued until ChatGPT created the reading passage and testing item that were structurally equivalent to those sampled from the CSAT (see the Appendix for the sample reading passages and testing items included in the instrument of the present study).

For each question type, we paired a human-generated item (i.e., the CSAT one) with a ChatGPT-generated item and asked participants to rate the characteristics and quality of the testing item of the same type generated by both methods. The testing components were presented in a randomized order of the two methods to prevent participants from depositing their sources. The survey was administered as a blind test.

Figure 1

Prompt Entered into ChatGPT to Create the Item Type of 'Identifying Changes in a Character's Mood'



Instrument

Four Likert-scale questionnaire items and one open-ended response item were given for each reading passage and testing item. Four questionnaire components were developed to measure participants' perceptions of (1) the naturalness of the writing flow, (2) the naturalness of the expressions, (3) the attractiveness of multiple-choice options (i.e., the extent to which distractions play a role in causing the difficulty of the items), and (4) the overall completion level of the testing item (i.e., the quality of the testing items in terms of their relevance to the target passage and whether the options are clearly written and homogenous in content), all on a 5-point Likert scale. For RQ2, in addition to the overall completion level (that purports to measure the overall quality), we also included a questionnaire item related to the attractiveness of multiple-choice options since our piloting showed that multiple options generated by ChatGPT are often not plausible. After presenting participants with the open-ended item, they were asked to elaborate upon the rationale for choosing a particular scale, if any. Figure 2 illustrates one of the reading passages and testing items, and the corresponding questionnaire items. The participants were asked first to read the passage and the testing item (on the left side of Figure 2) and then complete the questionnaire items (on the right side of Figure 2).

Figure 2

The Example of the Passage, the Testing Item, and the Questionnaire Items

Which of the following is consistent with the below announcement about a Charity Walk Event?

Charity Walk Event

Join a charity walk hosted by the Riverfront Park! This event supports the local animal shelter.

-**When & Where:** Sunday, May 15, 9:00 a.m./Riverfront Park.

-**How to Join the Walk:** Individual or team registration is accepted.

-Pay your registration fee of $20 as a donation.

-**Activities:** Walk a 5K route along the riverfront.

-With an additional $10 donation, you can participate in a pet adoption fair.

※Water and snacks will be provided.

Click here to register now!

① The event is held on a weekday.

② The event is held at the Riverside Park.

③ The event is free to participate.

④ The event is for abandoned animals.

⑤ The event includes a silent auction.

1. Determine the naturalness of the flow of this passage (1 = very unnatural ~ 5 = very natural).

1	2	3	4	5

2. Determine the naturalness of the English expressions of this passage (1 = very unnatural ~ 5 = very natural).

1	2	3	4	5

3. Determine the attractiveness of multiple-choice options for this question (1 = not attractive at all ~ 5 = very attractive).

1	2	3	4	5

4. Determine the overall completion level of the testing item (1 = lowest quality ~ 5 = highest quality).

1	2	3	4	5

5. If you would like to elaborate on the rationale for choosing a particular scale for one of the questionnaire items above, please do so.

_____

Data Analysis

For the data analysis, the participants' responses to the Likert-scale questionnaire items were first entered into a SPSS worksheet, with their responses to the CSAT-sampled reading passages and testing items and their ChatGPT-generated counterparts being coded differently. Next, their responses to each Likert-scale questionnaire item were averaged across the five reading passages and testing items extracted from the CSAT, and the same procedure was followed for the ChatGPT-generated passages and testing items. Then, four paired *t*-tests were conducted, respectively, with each Likert-scale item, with a Bonferroni correction adjusting for the alpha level (.05/4 = .0125).

The participants' responses to the open-ended questionnaire items were first distinguished between the ones given for the CSAT's reading passages and testing items, and those given for the ChatGPT ones. Afterward, they were grouped together according to theme (i.e., naturalness of flow, naturalness of English expressions, attractiveness of multiple-choice options, overall completion level of the testing item). The responses that were not relevant to any of the themes or did not offer a rationale for choosing a particular scale were excluded from the dataset.

Results

Table 1 shows the mean rating of the participants on the CSAT and ChatGPT-developed testing components, along with the results of paired *t*-tests.

Table 1

Participants' Mean Rating on the CSAT and ChatGPT-developed Testing Items and the Paired t-Test Results

Survey item	CSAT Mean (*SD*)			ChatGPT-developed Mean (*SD*)			*t*-value	95% confidence intervals	Cohen's *d*
	Pre-service	In-service	Total	Pre-service	In-service	Total			
Naturalness of flow	4.56 (.48)	3.98 (.61)	4.43 (.56)	4.40 (.40)	4.00 (.69)	4.31 (.50)	2.23	[.01, .22]	.24
Naturalness of English expressions	4.46 (.53)	4.11 (.52)	4.38 (.55)	4.36 (.53)	4.07 (.73)	4.30 (.59)	1.71	[–.01, .18]	.14
Attractiveness of multiple-choice options	4.35 (.47)	3.64 (.78)	4.19 (.63)	3.86 (.66)	3.29 (.79)	3.73 (.72)	6.71*	[.32, .60]	.64
Overall completion level of the testing item	4.39 (.43)	3.44 (.69)	4.18 (.64)	4.00 (.64)	3.44 (.79)	3.88 (.71)	3.40*	[.12, .47]	.42

Note. *p < .0125 (adjusted with a Bonferroni correction).

As Table 1 demonstrates, the pre-service group gave overall higher ratings than their in-service counterparts, regardless of the types of passages and testing items (i.e., the CSAT items or those developed by ChatGPT) or survey components. Regarding the CSAT, the results of the independent *t*-tests revealed that the pre-service group gave significantly higher ratings than the in-service group in terms of naturalness of flow (t = 3.28, p = .002) and overall completion level of the testing item (t = 4.38, p = .001), but not for the naturalness of English expressions (t = 1.92, p = .06) or the attractiveness of multiple-choice options (t = 2.90, p = .013), when a Bonferroni correction was adjusted for the alpha level (.05/4 = .0125). In the case of ChatGPT-developed testing components, the two groups' ratings were not significantly different, when a Bonferroni correction was adjusted for the alpha level, in terms of

naturalness of flow (t = 1.86, p = .087), naturalness of English expressions (t = 1.45, p = .15), attractiveness of multiple-choice options (t = 2.40, p = .02), and overall completion level of the testing item (t = 2.47, p = .017).

To answer the first research question, the participants rated the naturalness of flow and English expressions of the target reading passages highly (over 4.3), with no significant difference being found between the CSAT and ChatGPT-developed items. Some participants gave open-ended responses with regard to the naturalness of the flow and expressions of the passages developed by ChatGPT, as follows:

The flow of this passage [the fourth ChatGPT-developed reading passage] *seems to be natural*. (Pre-service teacher #35)

The sentences in this passage [the third ChatGPT-developed reading passage] *flow well overall*. (Pre-service teacher #13)

The flow, expression, and composition of this passage [the first ChatGPT-developed reading passage] *seem appropriate*. (In-service teacher #1)

As for the second research question, the CSAT items (M = 4.19) were rated significantly higher than the ChatGPT-based ones (M = 3.73) (p < .0125, adjusted with a Bonferroni correction) in terms of the attractiveness of multiple-choice options. Indeed, one of the most frequent comments regarding the opened-ended responses for the ChatGPT-based items concerned the lack of attractive distractors (n = 22). Some examples are given below:

There are no compelling option choices for this question [the question for the first ChatGPT-developed reading passage]. (Pre-service teacher #16)

Some options in this question [the question for the second ChatGPT-developed reading passage] *do not make sense at all.* (Pre-service teacher #30)

A significant modification is required for this question [the question for the second ChatGPT-developed reading passage]. *For example, the phrase in the fifth option is not even mentioned in the passage*. (In-service teacher #10)

The overall attractiveness of the options [for the question for the third ChatGPT-developed reading passage] *seems to be diminishing.* (In-service teacher #9)

The first option [in the question for the fifth ChatGPT-developed reading passage] *could also be the answer, I believe*. (In-service teacher #7)

As seen from the comments, some of the incorrect options generated by ChatGPT were deemed rather unattractive, or even as having the potential to be considered a correct answer. Although there were some negative comments about the options' attractiveness regarding the CSAT items as well, they were rare (n = 3).

For the other questionnaire item regarding the second research question, it was found that participants rated the completion level of the CSAT items (including the reading passages and testing components) (M = 4.18) significantly higher than the ChatGPT-developed ones (M = 3.88) (p < .0125, adjusted with a Bonferroni correction). Some of the relevant comments for the CSAT items are given below:

I think this testing item [for the third CSAT reading passage] *can be solved only when you fully understand the text and analyze the results accurately*. (Pre-service teacher #17)

The content of the [first CSAT] *passage is easy, but it is designed such that test takers should read both the passage and the options carefully in order to choose the correct answer*. (In-service teacher #3)

Considering the logical flow, it seems that there are a lot of conjunctive adverbs in the [fourth CSAT] *passage. However, the overall completion of the testing item is high*. (In-service teacher #2)

To summarize, the CSAT and ChatGPT-developed reading passages were identified as similar in terms of naturalness of the target passages' flow and expressions. In contrast, the former was judged as including more attractive multiple-choice options and as having a higher completion level regarding the testing items.

Discussion and Conclusion

In the present article, we have examined whether ChatGPT could generate L2 reading passages and testing items on par with those created by human experts. To this end, we administered a blind test with the pre- and in-service English teachers in the South Korean EFL context. Our findings revealed that ChatGPT is indeed capable of generating L2 reading passages that have a similar level of naturalness in terms of flow and expressions as those written and developed by human experts, as perceived by both participant groups. This outcome is consistent with recent studies (e.g., Ahn, 2023; Kwon & Lee, 2023), which have demonstrated that ChatGPT has a remarkable ability to solve L2 reading comprehension tasks. Notably, the present study further revealed ChatGPT's potential for designing L2 reading comprehension tests. We have also noted that, given the results related to the significant difference between the CSAT and ChatGPT-developed testing items (in terms of the participants' perception of the attractiveness of multiple-choice options and overall level of completion of the testing item), human teachers would indeed still be important for revising the testing items developed by ChatGPT, as seen in prior research (Shin, 2023). As evidenced in several excerpts from the participants' open-ended responses, there seems to be much room for improvement in ChatGPT's ability to construct well-designed testing components.

Given this study's results, our tentative answer to the question of whether ChatGPT is needed in L2 teaching was positive. That is, ChatGPT could help EFL teachers to generate passages for L2 reading and testing items more conveniently, and in a very short time period, thereby significantly reducing their workload. Meanwhile, teacher involvement in revising the generated testing items seems crucial, at least given the current state of ChatGPT's capacity. Thus, while it is not yet possible to completely hand over L2 testing item creation to ChatGPT, EFL teachers are strongly encouraged to explore its potential, given its powerful ability to assist them.

The following procedures could be followed by EFL teachers and practitioners who wish to create testing items using ChatGPT. First, they should identify a set of target skills and abilities (e.g., inferring a target passage's main idea) of interest in the given domain (e.g., reading) and examine the type of testing items that purport to assess each skill and ability. Next, they should collect model testing items from extant L2 tests that have high reliability and validity to examine whether the selected testing items fit their purposes (i.e., measuring the target skills and abilities). Then, they should draft the prompt (e.g., generate a new passage on a different topic and a multiple-choice question with 5 choices, as follows: [A model reading passage and testing item]) with which to generate each type of testing item, and revise it if the output is not satisfactory.

Finally, we provide the following suggestions for future research. First, a larger scale study with EFL learners, rather than pre- and in-service teachers, is needed to examine this same issue from learners' perspectives. Second, researchers are encouraged to use ChatGPT for generating testing items in other linguistic dimensions (e.g., listening, grammar) and to examine its potential in such areas. Lastly, a longitudinal study on L2 teachers' engagement in developing language testing components with ChatGPT and its washback effect would be expected to contribute to the growth of the current research branch on using ChatGPT in language teaching and learning.

Acknowledgements

We thank the pre- and in-service teachers who participated in the current study. We also appreciate the reviewers for their insightful feedback.

References

Ahn, Y. Y. (2023). Performance of ChatGPT 3.5 on CSAT: Its potential as a language learning and assessment tool. *Journal of the Korea English Education Society*, *22*(2), 119–145.

Al-Yahya, M. (2014). Ontology-based multiple choice question generation. *The Scientific World Journal*, *3*, 274949. https://doi.org/10.1155/2014/274949

Bormuth, J. (1969). *On a theory of achievement test items.* University of Chicago Press.

Chan, C. K. Y., & Tsi, L. H. Y. (2023). The AI revolution in education: Will AI replace or assist teachers in higher education? *arXiv*. https://doi.org/10.48550/arXiv.2305.01185

Ghumra, F. (2022, March). OpenAI GPT-3, the most powerful language model: An overview. *e-Infochips*. https://www.einfochips.com/blog/openai-gpt-3-the-most-powerful-language-model-an-overview/

Kim, T. (2023). *Can ChatGPT be an innovative tool?: The use cases and prospects of ChatGPT* (NIA_The AI Report 2023-1). NIA AI Future Strategy Center.

Klein, S., Aeschlimann, J. F., Balsiger, D. F., Converse, S. L., Court, C., Foster, M., Lao, R., Oakley, J. D., & Smith, J. (1973). *Automatic novel writing: A status report* (Technical Report 186). The University of Wisconsin, Computer Science Department.

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

Korea Institute for Curriculum and Evaluation. (2019). *2020 College Scholastic Ability Test: English section*. Korea Institute for Curriculum and Evaluation. https://www.suneung.re.kr/boardCnts/fileDown.do?fileSeq=b8cc879f115f67b90ace7b59c57641a8

Kumar, V., Boorla, K., Meena, Y., Ramakrishnan, G., & Li, Y. F. (2018). Automating reading comprehension by generating question and answer pairs. In D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji & L. Rashidi. (Eds.) *Advances in knowledge discovery and data mining* (pp. 335–348). Springer International Publishing.

Kwon, S-K., & Lee, Y. T. (2023). Investigating the performance of generative AI ChatGPT's reading comprehension ability. *Journal of the Korea English Education Society*, *22*(2), 147–172.

Lee, J. H., Shin, D., & Noh, W. (2023). Artificial intelligence-based content generator technology for young English-as-a-foreign-language learners' reading enjoyment. *RELC Journal*. https://doi.org/10.1177/00336882231165060

Liu, M., Calvo, R. A., & Rus, V. (2012). G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue and Discourse*, *3*(2), 101–124. https://doi.org/10.5087/dad.2012.205

Meehan, J. R. (1977). TALE-SPIN: An interactive program that writes stories. In R. Reddy (Ed.), *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (pp. 91–98). Morgan Kaufmann Inc.

Shin, D. (2023). A case study on English test item development training for secondary school teachers using AI tools: Focusing on ChatGPT. *Language Research*, *59*(1), 21–42. https://doi.org/10.30961/lr.2023.59.1.21

Tao, H. B., Diaz, V. R., & Guerra, Y. M. (2019). Artificial intelligence and education: Challenges and disadvantages for the teacher. *Arctic Journal*, *72*(12), 30–50.

von Davier, M. (2018). Automated item generation with Recurrent Neural Networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Appendix. The Sample Reading Passages and Testing Items Included in the Blind Test

Reading passages and testing items 1-1, 2-1, 3-2 in the Appendix are copyrighted by the Korea Institute for Curriculum and Evaluation (2019).

1.1 (2019 CSAT English section Q#19)

Which of the following is the most appropriate for capturing Jonas' emotional change, as revealed in the following article?

Looking out the bus window, Jonas could not stay calm. He had been looking forward to this field trip. It was the first field trip for his history course. His history professor had recommended it to the class, and Jonas had signed up enthusiastically. He was the first to board the bus in the morning. The landscape looked fascinating as the bus headed to Alsace. Finally arriving in Alsace after three hours on the road, however, Jonas saw nothing but endless agricultural fields. The fields were vast, but hardly appealed to him. He had expected to see some old castles and historical monuments, but now he saw nothing like that awaiting him. "What can I learn from these boring fields?" Jonas said to himself with a sigh.

① excited → disappointed ② indifferent → thrilled
③ amazed → horrified ④ surprised → relieved
⑤ worried → confident

1-2 (Chat GPT)

Which of the following is the most appropriate in capturing Maria's emotional change, as revealed in the following article?

As she stepped onto the stage, Maria's heart began to race. She had always loved singing, but this was the first time she would be performing in front of a large audience. She had been practicing for weeks, preparing for this moment. As the music started playing and the lights came on, Maria felt a surge of excitement. Her voice was strong and clear, and the audience seemed to be enjoying her performance. But as she reached the chorus, Maria's confidence began to waver. She had missed a note, and now she was worried she would make another mistake. By the end of the song, Maria was feeling disappointed with her performance, despite the enthusiastic applause from the audience.

① excited → disheartened ② indifferent → thrilled
③ amazed → horrified ④ surprised → relieved
⑤ worried → confident

2-1 (2019 CSAT English section Q#28)

Which of the following is consistent with the below announcement about the 2019 Badminton Challenge for Charity?

2019 Badminton Challenge for Charity

Join the charity tournament event hosted by Cliffield Community Center! This event supports Salke Children's Hospital.

When & Where

·Saturday, November 23, 2:00 p.m.

·Cliffield Sports Center.

How to Join the Tournament

·Make a two-member team.

·Pay your team's $100 entry fee as a donation.

Activities

·Challenge last year's champion team to a 3-point match.

·With an additional $20 donation, you can learn badminton skills from professional players.

※ Rackets and shuttlecocks will be provided.

Click here to register now!

① It is hosted by Salke Children's Hospital.
② Three people should form a team and participate.
③ The participation fee is $100 per person.
④ You can learn badminton skills if you donate 20 dollars extra.
⑤ Racket and shuttlecock are not provided.

2-2 (Chat GPT)

Which of the following is consistent with the below announcement about a Charity Walk Event?

Charity Walk Event

Join a charity walk hosted by the Riverfront Park! This event supports the local animal shelter.

-**When & Where:** Sunday, May 15, 9:00 a.m./Riverfront Park.

-**How to Join the Walk:** Individual or team registration is accepted.

-Pay your registration fee of $20 as a donation.

-**Activities:** Walk a 5K route along the riverfront

·With an additional $10 donation, you can participate in a pet adoption fair.

※Water and snacks will be provided.

Click here to register now!

① The event is held on a weekday.
② The event is held at the Riverside Park.
③ The event is free to participate.
④ The event is for abandoned animals.
⑤ The event includes a silent auction.

3-1 (Chat GPT)

Choose the most appropriate sentence for the blank in the below passage.

The famous novel *To Kill a Mockingbird* by Harper Lee explores the issue of racial injustice in the southern United States. The main character, Scout Finch, grows up in the 1930s in the small town of Maycomb, Alabama, where she witnesses the prejudice and narrow-mindedness that exist in the town. _____ Through her experiences and the influence of her father, Atticus (a lawyer who defends an African American man accused of assaulting a white woman), Scout learns important lessons about tolerance, compassion, and the pursuit of justice.

① Scout also learns how to deal with her neighbors.
② The town is also struggling with the Great Depression.
③ Scout's curiosity and adventurous spirit often get her into trouble.
④ The trial of the African American man serves as a catalyst for Scout's growth.
⑤ Scout becomes friends with a mysterious neighbor named Boo Radley.

3-2 (2019 CSAT English section Q#32)

Choose the most appropriate sentence for the blank in the below passage.

The Swiss psychologist Jean Piaget frequently analyzed children's conception of time via their ability to compare or estimate the time taken by pairs of events. In a typical experiment, two toy cars were shown running synchronously on parallel tracks, _____. The children were then asked to judge whether the cars had run for the same time and to justify their judgment. Preschoolers and young school-age children confused temporal and spatial dimensions: starting times are judged by starting points, stopping times by stopping points, and durations by distance, although each of these errors does not necessitate the others. Hence, a child may claim that the cars started and stopped running together (correct) and that the car that stopped further ahead ran for more time (incorrect).

① one running faster and stopping further down the track
② both stopping at the same point further than expected
③ one keeping the same speed as the other to the end

④ both alternating their speed but arriving at the same end
⑤ both slowing their speed and reaching the identical spot

About the Authors

Dongkwang Shin received his PhD in Applied Linguistics from Victoria University of Wellington and is currently a Professor at Gwangju National University of Education, South Korea. His research interests include corpus linguistics, CALL, and AI-based language learning.

E-mail: sdhera@gmail.com

Jang Ho Lee received his DPhil in education from the University of Oxford, and is presently a Professor at Chung-Ang University, South Korea. His areas of interest are CALL, L1 use in L2 teaching, and vocabulary acquisition. All correspondence regarding this publication should be addressed to him.

E-mail: jangholee@cau.ac.kr