Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 104152 (2026)
https://doi.org/10.29140/lea.2026.104152

An Exploratory Comparison of ChatGPT-generated and Human-made Japanese Reading Comprehension Test Items

GILBERT DIZON
RYO KUROSE

Kansai University, Japan
gdizon@kansai-u.ac.jp

Kyoto University of Foreign Studies, Japan
kuryose@gmail.com

Abstract

While existing research involving generative artificial intelligence (GenAI) has focused primarily on English, little is known about the technology's impact on other languages, particularly in the context of second language (L2) assessment. Thus, this study examined the potential of GenAI to support non-English language assessment by comparing expert-developed reading comprehension passages and test questions and ChatGPT 4.0-generated test materials based on the top two levels (i.e., N1 and N2) of the Japanese-Language Proficiency Test (JLPT). Japanese university language instructors (N = 16) blindly evaluated both sets of materials in terms of naturalness of the writing flow, naturalness of the Japanese expressions, attractiveness of the multiple-choice options, and overall completion level of the testing item. Descriptive statistics and Mann-Whitney U tests revealed no significant differences between the two types of reading assessment materials. Qualitative survey responses indicated some concerns regarding the naturalness of the expressions, cohesion, and answer choice quality, but no major perceived differences overall. These findings suggest that GenAI may serve as a useful supplementary tool for developing JLPT preparation materials, potentially reducing time and cost burdens. However, human expertise remains necessary to refine and ensure the quality of GenAI-generated assessment materials.

Keywords: Generative artificial intelligence, Japanese as a second language, Japanese language proficiency test, languages other than English

Data Availability Statement: *All relevant data are within this paper.*

Introduction

The public release of ChatGPT in late 2022 sparked significant interest in generative artificial intelligence (GenAI) within second language (L2) education. In a recent systematic review of GenAI-related language learning literature from 2023 to 2024, Li et al. (2025) analyzed 144 peer-reviewed studies. Among this body of work, assessment was one of the most frequently studied areas, particularly in terms of L2 writing. Findings from these studies indicate that GenAI is a useful tool for the automated assessment of L2 writing (Mizumoto & Eguchi, 2023; Pack et al., 2024). Nonetheless, research is limited when it comes to other aspects of L2 assessment beyond writing.

Another important finding from Li et al. (2025) is the paucity of research involving languages other than English (LOTE). In other words, English-focused research dominated, with 86% of the included studies examining English language learning. This finding mirrors the trends identified in other major systematic reviews on GenAI and language education (Lee et al., 2025; Wang et al., 2025). This language imbalance limits the generalizability of the findings and hinders our understanding of GenAI across diverse language-learning settings as the technology's performance may be dependent on the language (Swinehart et al., 2025). Consequently, it is critical to investigate non-English L2s to better understand GenAI's impact across diverse contexts.

One potential avenue to address this imbalance is the context of the Japanese-Language Proficiency Test (JLPT), a standardized Japanese language assessment developed jointly by The Japan Foundation and Japan Educational Exchanges and Services. Scores on the JLPT often inform decisions on education and employment, with the top two levels (N1 and N2) largely considered necessary for university study in Japan (Nishizawa et al., 2022). As global interest in the Japanese language and culture has increased, so has the number of JLPT test takers. More than 1.4 million individuals took the JLPT in 2024, which is the highest number in a single year to date (The Japan Foundation / Japan Educational Exchanges and Services, n.d.). Despite the significance of the JLPT, no study has explored whether GenAI can be used to develop quality test preparation materials for the test. To address this gap, the goal of this study was to compare language instructors' perceptions of human-developed Japanese reading comprehension test items with those generated by ChatGPT.

Literature Review

L2 Assessment with GenAI

Much of the literature involving GenAI for L2 assessment has focused on automated essay scoring (AES). Mizumoto and Eguchi (2023) were among the first researchers to evaluate the reliability and accuracy of GenAI in this context. They used ChatGPT 3.5 to automatically score over 12,000 Test of English as a Foreign Language (TOEFL) essays. Results demonstrated that ChatGPT produced similar benchmark scoring to the professional human ratings, thereby highlighting GenAI as a viable alternative to human raters when it comes to L2 essay scoring. Later studies using different language learning models (LLMs), AI-based systems that use transformer architecture and large datasets to generate accurate and human-like language (Yang & Chen, 2025), and essay corpora have resulted in similar findings. In other words, GenAI exhibits acceptable levels of reliability when used for AES (Pack et al., 2024; Yamashita, 2024). Specifically, varying reliability scores between 0.59 and 0.92 were found in Pack et al., depending on the specific LLM involved. In Yamashita, quadratic weighted kappa (QWK) was used to assess the agreement between ChatGPT 4.0 and human raters. QWK values between 0.57 and 0.74 were found, which indicates a moderate level of agreement. In short, results from Yamashita's research suggest that ChatGPT's ratings were consistent and comparable to those of

human raters. Additionally, finetuning LLMs (i.e., training LLMs for a specific task) has been shown to further improve AES accuracy and reliability (Wang & Gayed, 2024). Taken together, these studies and others (e.g., Pfau et al., 2023; Uchida & Negishi, 2025) illustrate that GenAI appears to be a reliable tool for L2 writing assessment.

However, these studies also raise some concerns when it comes to using GenAI to assess L2 writing. For instance, Pack et al. (2024) noted that some of the LLMs in their study sometimes demonstrated peculiar behaviors. More specifically, some LLMs would assign scores on a continuous scale even though the rubric was ordinal. Also, one of the LLMs would sometimes complete students' unfinished paragraphs, while also including this LLM-generated writing when scoring, thus artificially inflating scores. These types of issues were also exhibited in the study by Wang and Gayed (2024). For example, the finetuned models tended to overrate lower-rated essays and rarely assigned 3 as a score based on a 5-point scale. Because of these limitations, many studies suggest (e.g., Mizumoto & Eguchi, 2023) that GenAI is best suited as a supplement to human raters instead of as a replacement.

Although findings on the use of GenAI for L2 writing assessment are promising, relatively fewer studies have examined the use of GenAI for assessment in non-writing contexts. One exception in the context of L2 reading is Shin and Lee (2023), who investigated whether ChatGPT could make L2 English reading comprehension test items that were comparable with those made by human experts. The results of their study, which involved pre- and in-service teachers in South Korea, indicated that the participants viewed the human-made and ChatGPT-generated reading test materials as similar in regard to the naturalness and attractiveness of the reading passages and overall quality of the test items or questions. A more recent study by Lin and Chen (2024) had similar findings. Specifically, the results of psychometric tests, a survey, and interviews indicated that the ChatGPT-generated L2 English reading comprehension test items were acceptable when compared to human-developed items. An additional study by Shin et al. (2025) investigated the perceptions that South Korean pre-service teachers had toward two AI-based L2 English reading test generators. While both test generators were viewed favorably, the one based on GPT-4 was better received. Other areas related to L2 assessment, including listening assessment (Stewart et al., 2025) and speaking assessment (Liu et al., 2025), have also resulted in positive outcomes. Collectively, these studies demonstrate that GenAI has great potential for both language teachers and learners in terms of language assessment, as the emerging technology can serve to produce reliable testing materials in L2 English. However, these findings may not apply to non-English languages as LLMs may lack sufficient training data to generate high-quality output in those languages (Swinehart et al., 2025). Therefore, research is needed to better understand the use of GenAI for assessment in non-English L2 contexts.

Generative AI and LOTEs

Although the majority of studies on GenAI in the L2 learning context have focused on English (Li et al., 2025), initial research on the use of LLMs for the teaching and learning of LOTEs sheds light on the benefits and limitations of the technology. In a study on the L2 Chinese context, Yang and Chen (2025) evaluated ChatGPT's ability to provide automated corrective feedback. Results from their study showed that ChatGPT 4.0 could reliably and accurately correct L2 Chinese grammatical errors with minimal over-correction. Conversely, a study by Guo (2024) on the L2 German context resulted in less positive findings. Namely, the L2 German participants in the study exhibited skepticism toward ChatGPT, highlighted by the low acceptance rate (65%) of the corrective feedback generated by the GenAI tool. Notwithstanding these negative results, Guo emphasized the ability of ChatGPT to provide instantaneous feedback, which is particularly relevant to L2 German teachers and learners due to the scarcity of resources in the language.

Other studies involving non-English languages demonstrate the varying performance of GenAI across different contexts. For example, Tran and Stell (2024) explored ChatGPT's capacity to recognize and generate different Mandarin and Vietnamese dialects. Findings revealed that ChatGPT was able to generate language in these dialects, but the reliability of its output was limited. For instance, many Vietnamese words in particular dialects were replaced with standard forms, and ChatGPT also tended to provide inaccurate factual information about these minority languages. In a novel study, Swinehart et al. (2025) developed a protocol to measure GenAI's performance across multiple languages. Using this protocol, the researchers compared Microsoft Copilot's responses to language-teaching prompts across 26 languages, with the output being judged by 46 university language instructors. Results indicated that while the GenAI chatbot's performance was superior in higher-resource languages (e.g., German, Japanese, Mandarin) compared to lower-resource languages (e.g., Polish, Turkish, Czech), instructor evaluations showed that Copilot's output was still valuable. In other words, teachers of lower-resource languages still found the GenAI tool to be useful for various tasks such as creating comprehension questions, summarizing text, and receiving writing feedback despite issues with output quality. A survey study by Dizon (2026) investigating ChatGPT's potential as a resource for self-directed foreign language learning further illustrates the potential disparity in perceptions and experiences between learners when utilizing GenAI for language learning. Specifically, learners of more commonly taught languages (MCTLs), namely, English, German, Spanish, and French, had significantly more positive perceptions of the usefulness of ChatGPT than learners of less commonly taught languages (LCTLs) or all other L2s. Qualitative findings from the study help explain this discrepancy, as many of the participants expressed doubts regarding the accuracy of the tool's output from both a language and factual perspective. In short, these studies underscore the variability of GenAI when it comes to output in different languages, as well as the mixed results in terms of language learners' and teachers' perceptions. Therefore, more studies are needed to investigate how GenAI can be used for specific purposes, such as language assessment, to gain a deeper understanding of how to utilize GenAI across diverse language contexts.

Research Questions

The aforementioned research illustrates how GenAI has been used for L2 assessment. The literature review also sheds light on the usability and performance of GenAI tools for the learning of LOTEs. Despite the valuable insights gained from these studies, several research gaps still ought to be addressed. First and foremost, research is limited when it comes to GenAI and the learning of non-English languages (Li et al., 2025). In fact, none of the 144 studies included in Li et al.'s review of GenAI and language learning research focused on L2 Japanese language assessment. Therefore, further research is needed to better understand GenAI's impact on the learning of LOTEs, such as Japanese. Furthermore, GenAI research on L2 assessment has primarily focused on L2 writing, while other areas related to L2 assessment have received relatively less attention. In light of these research gaps, the following research question was addressed in this study:

- How do university language teachers in Japan perceive the JLPT and ChatGPT-generated reading passages and test items?

Methodology

Participants

Snowball sampling was used to recruit participants for the study. In this form of convenience sampling, researchers utilize their social networks to establish initial connections with participants, which then expand as those participants recruit additional individuals, leading to an increase in participants

(Parker et al., 2019). According to Parker et al., this type of sampling is ideal when recruiting from populations that are difficult to reach, which fits the description of the participants in the present study.

A total of 37 university language instructors participated in the survey-based study. All of the participants were native (L1) Japanese or advanced L2 Japanese speakers. Sixteen of the instructors fully completed the survey, so only data from these participants were analyzed in the study. Table 1 displays the demographics of the participating language instructors. In summary, female participants outnumbered male participants, and their ages were relatively evenly distributed across the different groups. The participants were primarily experienced university language instructors with more than 11 years of teaching experience.

Table 1. *Participant demographic information*

Gender identity	Number
Female	10
Male	6
Age	
30–39	1
40–49	5
50–59	5
60+	5
Years of language teaching experience	
1–5	2
6–10	1
11–20	8
20+	5

Research Instrument

The Japanese-language survey was adapted from Shin and Lee (2023) and consisted of two sections. The first section informed respondents of the general aims of the study and included one question related to consent. Similar to Shin and Lee, participants were not explicitly told that the survey would contain JLPT and ChatGPT-generated reading passages. Instead, they were informed that the study was related to the topic of AI and Japanese language proficiency. If the respondents agreed to take part in the study, they moved on to the second section, which comprised the four reading passages (i.e., two JLPT and two ChatGPT-generated) and corresponding reading comprehension test items.

Two JLPT reading passages, taken from the two highest levels of the exam (i.e., N1 and N2), were selected from sample test materials on the JLPT website (The Japan Foundation / Japan Educational Exchanges and Services, n.d.). Using a similar prompt as Shin and Lee (2023), ChatGPT 4.0 was used to generate two additional reading passages and corresponding comprehension questions based on the selected JLPT materials. In total, two ChatGPT-generated reading passages and two test items were included in the study (see Appendices 1 and 2 for sample JLPT and ChatGPT-generated test materials and their English translations). The exact prompt used was: *Draft a new passage with a different topic with a similar multiple-choice question, as follows*. Figure 1 depicts the prompt along with one of the JLPT passages used in the study.

Figure 1. *Prompt used and example ChatGPT-generated reading passage and test item.*

For each reading passage and corresponding test question, the participants were asked to evaluate the materials based on four criteria: (1) naturalness of the writing flow; (2) naturalness of the expressions; (3) attractiveness of the multiple-choice options (i.e., the extent to which distractors contribute to item difficulty); and (4) overall level of completion of the testing item (i.e., relevance to the target passage and the clarity and homogeneity of the answer options). These four components were adopted as they were the same ones used in Shin and Lee (2023). For each criterion, the respondents provided ratings using a sliding scale ranging from 1 (e.g., very unnatural) to 5 (e.g., very natural) in 0.1 increments (e.g., 2.1, 2.2, 2.3), thus allowing for more fine-grained evaluations. In total, the scale included 41 points. The Cronbach's alpha of the survey was 0.85, thus indicating a high level of internal consistency. The participants completed the survey blind. In other words, they were not informed whether the reading passages and test items had been created by humans or generated using GenAI. One open-ended question followed each reading passage and item, thus allowing the respondents to elaborate on their ratings. The prompt was: "If you would like to elaborate on the rationale for choosing a particular response for one of the survey items, please do so here."

Data Collection

Data were collected via an anonymous online survey using the researchers' professional networks and an online community focused on language teaching with technology. Initially, data collection took place during the Fall 2024 semester and targeted university instructors teaching Japanese as a second language (JSL) in Japan. However, due to the small number of fully completed responses at that time (n = 3), the recruitment criteria were expanded during the Spring 2025 semester to include all university language instructors who self-reported as either L1 Japanese speakers or advanced L2 Japanese speakers (i.e., CEFR C2), regardless of the language they taught. This self-report approach was adopted to allow for broader participation and inclusivity. Thirteen fully completed responses were collected at this time, which brought the total number of valid responses to 16.

Data Analysis

Descriptive statistics, i.e., mean (M) and standard deviation (SD), were calculated with Excel to assess the participants' opinions of the JLPT and ChatGPT-generated reading comprehension test items. Due to the small sample size, the non-parametric Mann-Whitney U test was used to determine if there were significant differences in the participants' perceptions between the two types of tests (i.e., JLPT vs. ChatGPT-generated). The authors coded the qualitative data using thematic analysis (Braun & Clarke, 2021) to provide deeper insights into their perceptions of the different reading passages and test items. Figure 2 shows an example of the researchers' coding process involving a sample from the qualitative survey data. Rather than using a neopositivist approach to thematic analysis, the reflective approach advocated by Braun and Clarke was utilized. Therefore, inter-coder reliability, an often-reported feature of qualitative research, was not measured.

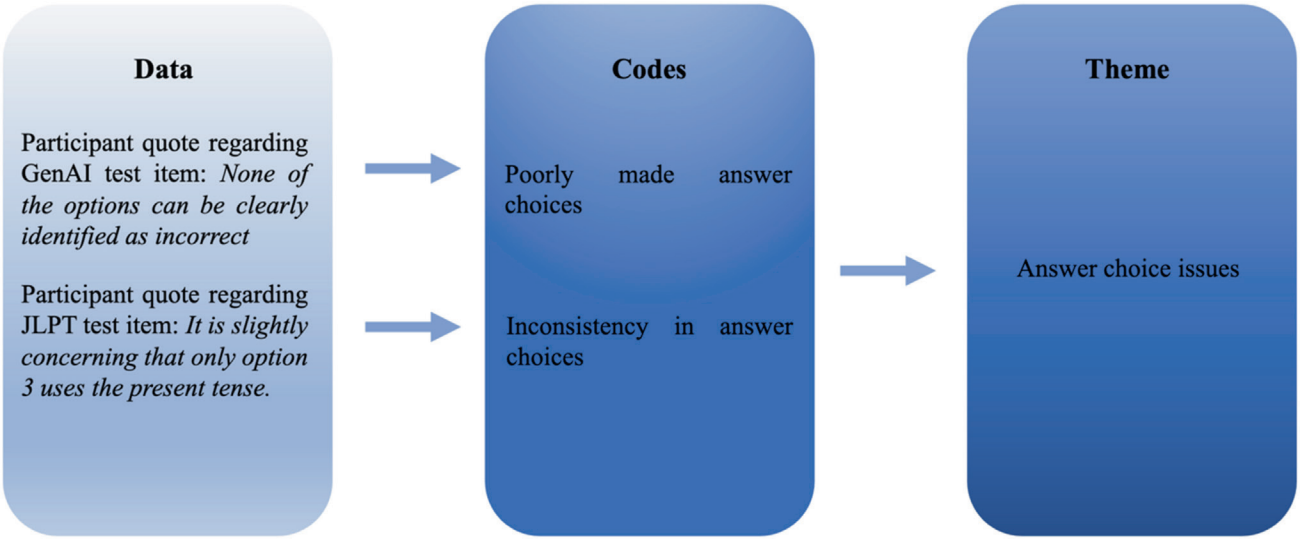

Figure 2. *Example of the coding process.*

Results

Table 2 shows the quantitative survey results comparing two JLPT and two ChatGPT-generated reading passages and test items. The mean values for the JLPT and ChatGPT-generated test materials indicate that the participants had moderately positive perceptions overall on a sliding scale ranging from 1.0 to 5.0. The mean scores of the tests were comparable, with the ChatGPT-generated reading passages and test items being viewed slightly more positively than the JLPT ones on three out of the four criteria. Specifically, ChatGPT received higher ratings for the naturalness of the Japanese expressions (M = 4.22, SD = 0.89), attractiveness of the multiple-choice options (M = 3.47, SD = 1.32), and overall completion level of the testing item (M = 3.58, SD = 1.29) than the JLPT (M = 4.07, SD = 0.94; M = 3.37, SD = 1.28; and M = 3.54, SD = 1.29 respectively). Only in terms of the naturalness of the writing flow did the JLPT receive a higher mean score (JLPT: M = 4.46, SD = 0.73; ChatGPT: M = 3.94, SD = 1.19).

In terms of naturalness of the writing flow, results from the Mann-Whitney U test showed that the difference between the JLPT and ChatGPT-generated reading passages was not significant, U = 404, p = 0.14. Non-significant results were also found regarding the naturalness of the Japanese expressions, U = 456, p = 0.45. When comparing the attractiveness of the multiple-choice options, no significant difference was found between the human-developed and GenAI-generated reading test questions,

U = 492.5, $p = 0.79$. Lastly, when it came to the overall completion level of the testing item, no significant difference was found, U = 492, $p = 0.79$. In summary, these results indicate that the participants perceived the human-made and GenAI-generated reading test materials to be of similar naturalness, attractiveness, and overall quality.

Table 2. *Quantitative survey results comparing JLPT and ChatGPT-generated reading passages and test items*

Survey construct	JLPT		ChatGPT	
	M	SD	M	SD
Naturalness of the writing flow	4.46	0.73	3.94	1.19
Naturalness of the Japanese expressions	4.07	0.94	4.22	0.89
Attractiveness of the multiple-choice options	3.37	1.28	3.47	1.32
Overall completion level of the testing item	3.54	1.24	3.58	1.29

Three themes were identified by the researchers from the thematic analysis. The first theme relates to perceived structural incoherence in the ChatGPT-generated readings. Specifically, some of the participants thought that the ChatGPT-generated readings did not have a clear, logical flow, which interfered with the readability of the passages. For instance, when discussing the first GenAI-developed readings, participants stated, "It is incoherent, and it's unclear what the main point is," and "Perhaps this is due to the author's style, but there were places that I could not understand without rereading several times." Similarly, another respondent raised concerns about coherence when evaluating the other ChatGPT-generated passage:

> Unnatural flow of the text: While paragraphs 1 to 3 focus on the idea that "modest actions such as garbage separation, recycling, and energy conservation are important," the final paragraph shifts the focus with the phrase "even in economic activities…" This causes the overall message to become inconsistent.

Interestingly, another theme identified through thematic analysis was unnatural expressions in the JLPT readings. In relation to one of the JLPT readings, one respondent noted, "The flow of the argument is easy to follow, but there are some unnatural expressions in the Japanese." When explaining the rationale for their evaluation of a different JLPT reading, another participant said the following:

> Unnatural aspect of the Japanese: "When driving a car, you shouldn't hit people; when cooking, you shouldn't cut your hand with a knife. That's just common sense." The way the parts are connected feels rough.

The final theme identified by the authors involved issues with the answer choices for the reading passages. Unlike the previous two themes, this theme was reflected in the participants' responses for both the JLPT and ChatGPT-generated reading comprehension items. When discussing the GenAI-created answer choices, participants commented that "the distractors could have been devised more carefully" and "none of the options can be clearly identified as incorrect." Likewise, the participants stated that there were problems with the JLPT answer choices, noting that "content not mentioned in the passage was presented as an answer choice," and that "none of the options seem to be correct." Another participant remarked that one of the answer choices for the JLPT readings stood out from the rest:

"It is slightly concerning that only option 3 uses the present tense." These quotes illustrate that the participants were concerned with the quality of the answer choices of the JLPT and ChatGPT-generated reading passages.

Discussion

Results from the descriptive statistics show similar mean values for the JLPT and ChatGPT-generated reading comprehension test materials. The Mann-Whitney U test showed that there was no significant difference between the human-made and GenAI-developed test items. Therefore, it can be concluded that participating Japanese university language teachers perceived the JLPT and ChatGPT-generated reading comprehension test materials to be similar in terms of naturalness, attractiveness, and overall quality. These results are in line with recent studies involving the use of GenAI for L2 English reading comprehension assessment (Lin & Chen, 2024; Shin and Lee, 2023) and demonstrate the potential of this emerging technology for developing reading comprehension materials for LOTEs like Japanese.

The qualitative results based on the participants' written responses showed that they had some concerns with the reading comprehension test materials in terms of the naturalness of the expressions (JLPT only), lack of cohesion (ChatGPT only), and quality of the answer choices (JLPT and ChatGPT). Even though the JLPT test materials were developed by human experts, the participants still perceived issues with the test items. One possible explanation for this apparent discrepancy may relate to the nature of the study. Namely, while the participants did not know the exact aims of the research, and assessed the materials blindly, they were informed that the study was related to Japanese language assessment and AI. Because of this, the participants may have adopted a more critical evaluation stance overall, leading them to focus on perceived deficiencies in the presented test materials, regardless of origin. Nonetheless, both the quantitative and qualitative findings from the study indicate that there were no perceived major differences between the two types of reading comprehension test materials. As such, this study advances research on GenAI and assessment by highlighting the potential usefulness of these digital resources in non-English learning contexts. This is particularly important for LOTEs such as Japanese since learners of these languages may have limited resources compared to English learners (Guo, 2024).

Conclusion

This exploratory study compared ChatGPT-generated and human-made Japanese reading comprehension test items. The results indicated that the participants viewed both types of reading test materials as comparable. Based on these findings, the authors believe that GenAI can be used to supplement existing human-developed JLPT materials. This may help reduce the time and cost of creating and/or purchasing reading comprehension materials for Japanese language teachers and learners. However, as noted by Shin and Lee (2023), human experts remain essential for refining and strengthening GenAI-developed reading comprehension test materials to ensure they meet standards similar to those of human-made standardized assessments.

While the findings of this exploratory study shed light on the potential of GenAI to create assessment materials for LOTEs, several limitations must be addressed. First and foremost, the small sample size limits the generalizability of the findings. Thus, future research should involve a larger number of in-service language instructors. Pre-service teachers could also be included to increase the sample size. Additionally, only four passages were compared in this study, which represents a relatively small dataset. Therefore, future research should include a greater number of reading passages and test items to improve the robustness and generalizability of the findings. Moreover, while the study contributes to our theoretical understanding of GenAI as a resource for generating quality reading assessment

materials in Japanese, it may also hold potential for addressing other sections of the test. Investigating other language components such as language knowledge or listening comprehension, would help broaden the theoretical implications of GenAI beyond L2 Japanese reading. Lastly, this study only examined GenAI-developed JLPT test materials at the N1 and N2 levels. It would therefore be worthwhile to see if ChatGPT and other GenAI tools can generate test materials for the other levels of the test, which are designed for less advanced Japanese learners.

Declarations

Ethics Approval and Consent

Informed consent was obtained from all participants in this study. Participation was voluntary, and no personally identifying information was collected. Due to the nature of the study (i.e., adult participants, anonymous, survey-based data collection), the research was exempt from ethics review by the authors' institutions.

Generative AI Use

The authors used ChatGPT 5.0 to support the editing and polishing of the English language in the manuscript. The tool was also used to assist in the translation of the participants' Japanese responses into English. Generative AI was not used in any other way, such as formulating research ideas or analyzing the results. The authors reviewed and edited the content as needed and take full responsibility for the content of this manuscript.

References

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Dizon, G. (2026). ChatGPT as a tool for self-directed foreign language learning. *Innovation in Language Learning and Teaching*, *20*(2), 334–350. https://doi.org/10.1080/17501229.2024.2413406

Lee, S., Choe, H., Zou, D., & Jeon, J. (2025). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335–359. https://doi.org/10.1080/10494820.2025.2498537

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence,* Article 100445. https://doi.org/10.1016/j.caeai.2025.100445

Liu, X. J., Wang, J., & Zou, B. (2025). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes*, *75*, Article 101505. https://doi.org/10.1016/j.jeap.2025.101505

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nishizawa, H., Isbell, D. R., & Suzuki, Y. (2022). Review of the Japanese-language proficiency test. *Language Testing*, *39*(3), 494–503. https://doi.org/10.1177/02655322221080898

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

Parker, C., Scott, S., & Geddes, A. (2019). Snowball sampling. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. Williams (Eds.), *SAGE research methods foundations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526421036831710

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3), Article 100083. https://doi.org/10.1016/j.rmal.2023.100083

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

Shin, D., Lee, J. H., & Kim, K. (2025). An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal*. Advance online publication. https://doi.org/10.1177/00336882251326284

Stewart, J., Anthony, L., Batty, A. O., Nakamura, K., Nicklin, C., McLean, S., & Tomaru, K. (2025). Can we reliably score meaning recall vocabulary tests using AI? A comparison of human vs. AI scoring. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2481397

Swinehart, N., Nguyen, P., & Yeh, E. (2025). A protocol for evaluating AI chatbots' capabilities for low-resource language teachers. *Language Learning & Technology*, *29*(1), 1–21. https://doi.org/10.64152/10125/73661

The Japan Foundation/Japan Educational Exchanges and Services. (n.d.). *Japanese-Language Proficiency Test*. https://www.jlpt.jp/e/index.cgi

Tran, H., & Stell, A. (2024). Beyond borders or building new walls? The potential for Generative AI in recolonising the learning of Vietnamese dialects and Mandarin varieties. *Australian Review of Applied Linguistics*, *47*(3), 284–308. https://doi.org/10.1075/aral.24135.tra

Uchida, S., & Negishi, M. (2025). Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI. *Research Methods in Applied Linguistics*, *4*(2), Article 100199. https://doi.org/10.1016/j.rmal.2025.100199

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2024.2371395

Wang, Y., Zhang, T., Yao, L., & Seedhouse, P. (2025). A scoping review of empirical studies on generative artificial intelligence in language education. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2509759

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, *3*(3), Article 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yang, C. T. Y., & Chen, H. H. J. (2025). ChatGPT and L2 Chinese writing: evaluating the impact of model version and prompt language on automated corrective feedback. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2453205

Appendix 1

JLPT Reading Passage and Test Item

音楽や美術、彫刻などの芸術は、聴く人、観る人の批評によって育てられる。悪い
演奏をしたら、良くない作品を出品したら、その芸術家は次に表舞台に出る機会を失う。ところが、医師や看護婦が行なう医療の実践を評価できる患者は作られてこなかった。一般の人に知識を与えず、医療について評価できる患者がいない。だから医のアートが発展しないのである。これは患者にとってはもとより、医療者にとっても不幸なことではないか。

（日野原重明『いのちの言葉』による）

筆者がここで最も言いたいことは何か。

1. きちんとした医療の実践ができない医療者は表舞台から去るべきだ。
2. 医療者を不幸にしないために、一般の人も医療の知識を持つべきだ。
3. 医療も芸術と同様に、医療の良し悪しを評価できる患者が必要である。
4. 悪い演奏や良くない芸術作品は、病院における医療事故と同じである。

ChatGPT-generated Reading Passage and Test Item

企業の成功は、そのリーダーシップの質によって大きく左右される。優れたリーダーは、社員のモチベーションを高め、組織全体のパフォーマンスを向上させることができる。例えば、透明性のあるコミュニケーションや、公正な評価制度を導入することは、社員の信頼を獲得するために不可欠である。しかし、現実には、多くのリーダーが自己中心的な決定を行い、社員の意見を軽視することがある。その結果、社員の士気が低下し、企業全体の生産性が落ちることになる。リーダーシップの質が企業の命運(めいうん)を握ることを理解し、常に自己改善を目指す姿勢が求められるのではないだろうか。

（山田太郎『企業経営の本質』による）

筆者がここで最も言いたいことは何か。

1. リーダーは透明性のあるコミュニケーションを重視すべきだ。
2. 公正な評価制度が企業成功の鍵である。
3. 優れたリーダーシップが企業の成功に不可欠である。
4. 自己中心的なリーダーは企業にとって有害である。

Appendix 2

English Translation of JLPT Reading Passage and Test Item

Arts such as music, painting, and sculpture are cultivated through the criticism of those who listen and observe them. If a musician performs poorly or an artist exhibits an inferior work, that artist will lose the opportunity to appear on the public stage again. However, patients capable of evaluating the medical practices performed by doctors and nurses have not been cultivated. Ordinary people have not been given sufficient knowledge, and therefore there are no patients who can evaluate medical care. As a result, the art of medicine does not develop. Is this not unfortunate not only for patients but also for medical professionals?

(Based on Shigeaki Hinohara, Words of Life)

What is the main point the author wants to convey here?

1. Medical professionals who cannot properly practice medicine should withdraw from the public stage.
2. To prevent medical professionals from becoming unhappy, ordinary people should also acquire medical knowledge.
3. Like the arts, medicine also requires patients who are able to evaluate the quality of medical care.
4. Poor musical performances and inferior artworks are the same as medical accidents in hospitals.

English Translation of ChatGPT-generated Reading Passage and Test Item

The success of a company is greatly influenced by the quality of its leadership. Excellent leaders can raise employee motivation and improve the performance of the organization as a whole. For example, introducing transparent communication and fair evaluation systems is essential for gaining the trust of employees. However, in reality, many leaders make self-centered decisions and disregard the opinions of their employees. As a result, employee morale declines and the overall productivity of the company falls. Should we not recognize that the quality of leadership determines the fate of a company and continually strive for self-improvement?

(Based on Taro Yamada, The Essence of Corporate Management)

What is the main point the author wants to convey here?

1. Leaders should place importance on transparent communication.
2. A fair evaluation system is the key to corporate success.
3. Excellent leadership is essential for a company’s success.
4. Self-centered leaders are harmful to companies.