English Language & Literature Teaching, *Vol. 31,* No. 3 Fall 2025
https://doi.org/10.35828/etak.2025.31.3.85

Enhancing L2 Testing through AI-Assisted Automated Item Generation

Youn-Kyoung Lee
(Kyungpook National University)

Lee, Youn-Kyoung. (2025). Enhancing L2 testing through AI-assisted automated item generation. *English Language & Literature Teaching, 31*(3), 85-103.

While automated item generation has gained attention across various disciplines, its application in English language teaching (ELT) remains relatively new. AI-assisted item generation has evolved from template- and rule-based models to statistical and deep learning approaches. Transformer-based systems (e.g., GPT) now produce more authentic and context-sensitive items, though issues of validity, fairness, and psychometric rigor persist. This study offers a comprehensive overview of automated item generation for language educators, drawing on a review of existing literature. It outlines major approaches, synthesizes findings from ELT-focused studies, and introduces four AI-powered tools: ChatGPT-4, genQue, Perplexity AI, and Twee, with practical guidance for classroom use. Findings indicate that ChatGPT generates items through prompt-based adaptation and iterative refinement. At the same time, Perplexity AI integrates GPT-4 with real-time search to provide source-based responses, though its mechanisms remain unclear. genQue, designed for the Korean College Scholastic Ability Test (CSAT), creates structured passages and items at varying difficulty levels but requires complex inputs. In contrast, Twee offers an intuitive interface for generating diverse, pedagogically valid tasks, supported by corpus-based authenticity and cognitive validity. The study concludes by stressing the need for further empirical research and encouraging teachers to engage with this emerging technology.

[AI-generated test items/automated item generation/AI-assisted L2 testing/L2 testing]

I. INTRODUCTION

Although the conceptual foundation for employing computer programs to generate test items can be traced back over half a century, most notably to Bormuth's (1969) early work, it is only within the past decade that automated item generation has garnered sustained scholarly attention in the field of education (Circi et al., 2023). With rapid advances in artificial intelligence and natural language processing, interest in automated item generation has been reignited, particularly due to its potential to address long-standing challenges in large-scale assessment design and implementation.

One of the most compelling advantages of automated item generation is its ability to produce a high volume of test items in a relatively short period (Gierl et al., 2012). This capability is especially beneficial for educators, institutions, and testing agencies that are frequently required to develop extensive item banks but lack the human resources or time to do so using traditional, expert-driven approaches. In such contexts, automated item generation offers a scalable and efficient alternative to conventional item development processes.

Moreover, automated item generation can be strategically leveraged to create item variants that are structurally similar yet cognitively distinct, thereby requiring test-takers to engage with different reasoning processes to arrive at correct answers. This capacity may help to reduce the risk of item memorization or recall, which has been identified as a threat to test validity, particularly in high-stakes assessment settings (Pugh et al., 2020).

Despite these promising developments, the application of AI-powered automated item generation in the domain of language assessment remains in its infancy. While recent studies have begun to explore how AI-driven tools can be used in English language teaching (ELT) contexts (e.g., Kohnke et al., 2023), there remains a lack of comprehensive, practitioner-oriented resources that introduce the principles, tools, and pedagogical considerations associated with automated item generation.

In response to this gap, the present article aims to (1) provide a review of recent research on AI-based automated item generation in L2 education, (2) introduce key AI tools, including ChatGPT, that ELT practitioners can employ to generate testing items, and (3) analyze these tools and finally offer practical guidelines and suggestions for effectively integrating automated item generation into classroom-based assessment practices for L2 education researchers and instructors.

II. REVIEW OF LITERATURE

1. Classifying Approaches to AI-Assisted Item Generation

Over the twentieth and twenty-first centuries, efforts to automate the generation of test items have undergone significant evolution. This line of inquiry dates back to Bormuth's seminal work in 1969. Building on this foundation, Kurdi et al. (2020) proposed a widely accepted classification of automated item generation methodologies into three primary categories: template-based, rule-based, and statistical approaches.

Template-based approaches generate items by employing fixed, predefined templates or item models. These systems follow established structural patterns to produce test items, ensuring consistency and reliability in item formatting. However, their rigid design often limits the diversity of item types they can produce, making it difficult to adapt to novel item formats or incorporate complex linguistic variations.

Rule-based approaches represent a more flexible alternative. These methods utilize predefined linguistic rules and syntactic structures to annotate sentences and generate test items accordingly. By leveraging explicit grammatical and lexical properties, rule-based systems can offer slightly greater adaptability than template-based models. Nevertheless, they still face significant constraints when applied to unfamiliar item types or tasks requiring higher-order reasoning or abstraction.

In contrast, statistical approaches—particularly those leveraging advances in machine learning—have shown considerable promise. These methods derive item-generation rules from large datasets of existing test items, enabling the production of a broader range of item types and formats. Furthermore, statistical approaches often yield more natural and contextually appropriate language, surpassing the often formulaic nature of template- or rule-based systems.

Building upon the foundation laid by statistical models, recent advances in deep learning have introduced a paradigm shift in automated item generation. Rather than relying solely on hand-crafted features or manually annotated data, deep learning models—especially those employing neural network architectures—learn latent patterns directly from large corpora, offering greater scalability and linguistic sophistication. For instance, von Davier (2018) applied a recurrent neural network (RNN), specifically a long short-term memory (LSTM) model, to generate test items aimed at assessing non-cognitive traits such as emotions and attitudes. This marked an important departure from prior approaches, as it demonstrated the capacity of AI-based systems to capture more subtle, affective dimensions of assessment with minimal human intervention. In the most recent wave of developments, deep learning—particularly sequence-to-sequence and transformer-based architectures—has emerged not merely as an extension of traditional statistical methods but as a

transformative, standalone paradigm in automated item generation. One of the earliest notable contributions, von Davier's (2018) study in *Psychometrika*, demonstrated that a long short-term memory (LSTM) recurrent neural network, trained on large item repositories such as personality inventories, could generate non-cognitive assessment items whose psychometric properties (e.g., factor structure) aligned closely with human-authored counterparts. Building on this foundation, transformer-based models (e.g., GPT-2, GPT-3) have been fine-tuned for domain-specific item generation. For instance, von Davier (2019) applied a GPT-2 model to create medical certification case vignettes and distractors, producing plausible drafts that item-writers could refine. More recently, Bezirhan and von Davier (2023) leveraged OpenAI's GPT-3 to automatically generate reading passages matched to specified Lexile levels, achieving high coherence and readability with minimal human editing. These transformer-based systems, often used in a prompt-engineering, human-in-the-loop framework, exhibit significantly higher linguistic fluency and item diversity than previous template- or rule-based outputs, indicating that deep learning constitutes a distinct, generative advancement rather than a mere integration of older approaches.

More recently, the advent of large-scale transformer-based models—such as ChatGPT, developed by Radford et al. (2019)—has brought renewed attention to automated item generation in educational contexts. These models are trained on massive corpora and excel in integrating broader contextual information, thereby producing output that is more coherent, naturalistic, and pedagogically relevant. The use of ChatGPT and similar models for item generation represents a significant advancement in generating authentic, context-sensitive language testing materials.

For example, Attali et al. (2022) employed a transformer-based deep language model to develop an interactive reading task, demonstrating how such approaches can support the creation of complex, multi-stage assessment tasks. This work highlights the growing sophistication and potential of AI-powered automated item generation, suggesting a shift from generating isolated test items to constructing more holistic and dynamic assessment experiences.

Taken together, the evolution of AI-assisted item generation reflects a trajectory from rigid, form-driven approaches toward increasingly flexible, data-driven, and context-aware methodologies. Template- and rule-based systems offer control and reliability but are constrained in their adaptability and linguistic richness. Statistical and early machine learning approaches expanded the scope of item types and improved surface-level fluency, yet they often lacked semantic depth and required extensive feature engineering. The emergence of deep learning—particularly with transformer-based models—has significantly enhanced the capacity to generate more authentic, coherent, and cognitively engaging items. However, despite these advancements, critical challenges remain. Issues such as content

validity, construct representation, ethical concerns (e.g., bias and fairness), and the interpretability of AI-generated items are still underexplored. Moreover, while deep learning models excel in linguistic generation, their pedagogical alignment and psychometric soundness require closer scrutiny. As such, future research should move beyond technical feasibility and engage more robustly with educational theory, test design principles, and human-AI collaboration to ensure that automated item generation contributes meaningfully and responsibly to assessment practices.

2. Recent Studies on AI-Assisted Item Generation

Kohnke et al. (2023) stated that ELT teachers may use ChatGPT to generate reading comprehension questions. In their study, they demonstrated how ChatGPT could be employed to generate both comprehension and expansion questions, using *Alice's Adventures in Wonderland* by Lewis Carroll as a source text. Although the authors did not provide a systematic evaluation of the pedagogical advantages or limitations associated with such AI-generated items, their illustration nonetheless draws attention to the growing applicability of generative AI models in assessment design within the language education domain.

Luo et al. (2022) examined AI-generated English vocabulary items and human expert-defined items to compare their item difficulty and item discrimination. The findings revealed that human expert-designed items were easier but showed more discriminating power than AI-generated items. The authors also stressed that AI-generated and expert-developed items may tap into distinct constructs, with the former prompting bottom-up processing strategies and the latter eliciting reliance on rote memorization. Thus, the AI-generated items seem to assess the receptive vocabulary ability (e.g., using the context/sentence to make sense of vocabulary), whereas the expert-designed items assess the extent to which test takers can memorize the teachers' examples which include the targeted words. In the related vein, Ha and Nguyen (2024) investigated the integration of AI-driven assessment tools in the development of ELT examinations by conducting interviews with a diverse cohort of ELT practitioners. Their findings indicated that AI was regarded as a valuable resource due to its distinct advantages, including automatic item generation, adaptive testing, enhanced feedback delivery, quality assurance, and the creation of innovative test formats. However, several limitations were also identified, such as the linguistic complexity and subtlety inherent in language use, technical constraints, ethical and bias-related concerns, and the continued necessity of human oversight and validation.

Furthermore, Shin (2023) investigated the practical application of a range of AI-powered tools—including ChatGPT, CopyAI, and Google Translate—for developing English reading comprehension items tailored to the CSAT, Korea's national university entrance examination.

The study was conducted within the context of a secondary-level teacher training program, where ten in-service English teachers participated in item creation tasks under the researcher's supervision. The findings indicated that the teachers generally perceived AI-assisted item generation as both time-efficient and user-friendly. However, the author emphasized that crafting prompts with precision—particularly those tailored to specific item types such as mood identification or contextual inference—was essential for generating high-quality assessment items, thereby highlighting the necessity of teacher expertise in guiding AI output.

Building on these findings, Shin and Lee (2023) implemented a blind evaluation study involving fifty pre-service and in-service English teachers in South Korea. Participants were asked to assess reading passages and associated test items, some of which were extracted from actual CSAT materials and others generated by ChatGPT. The results revealed that there was no statistically significant difference in participants' perceptions regarding the linguistic naturalness and coherence of the reading passages. However, the CSAT items received significantly higher ratings in terms of the appeal and plausibility of the multiple-choice options, as well as the overall completeness of the items. These findings are in line with Gierl and Lai's (2013) earlier discussion regarding the challenges in ensuring the quality of distractors produced via automated item generation systems. In light of this, Shin and Lee (2023) recommended that English teachers exercise critical judgment and make necessary revisions when incorporating AI-generated test items into their instructional or evaluative practices.

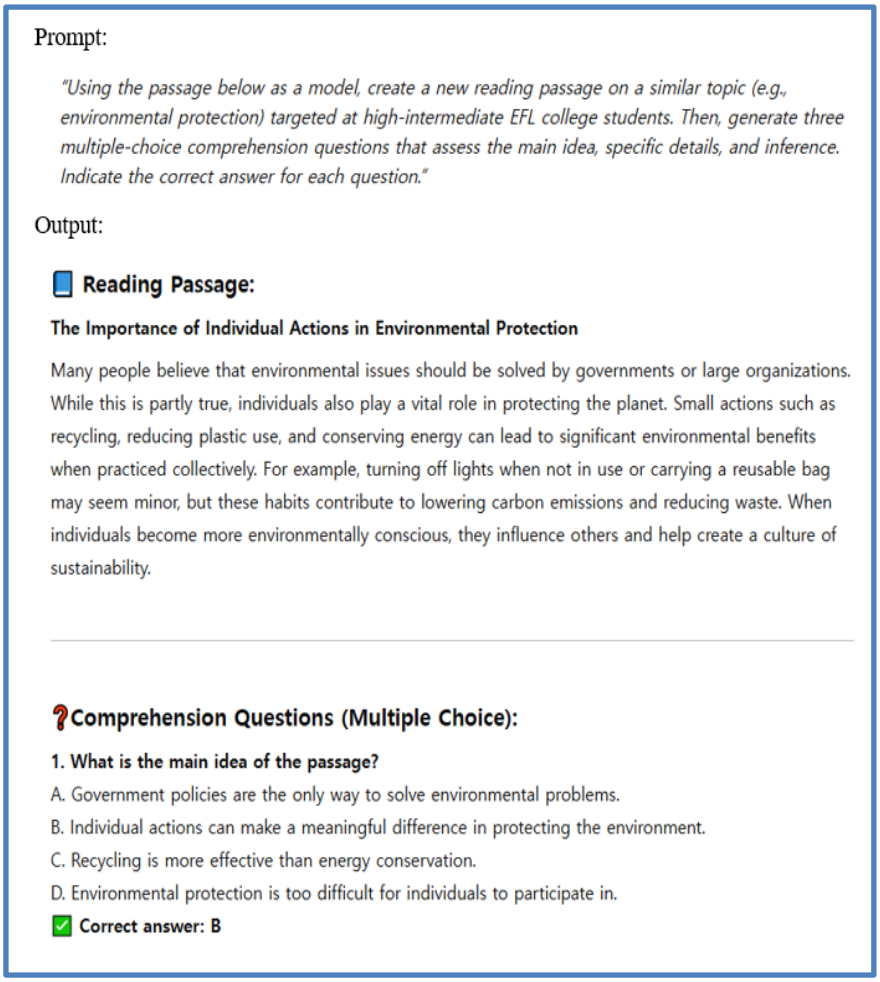
Taken together, recent studies underscore both the promise and the limitations of AI-assisted item generation in the field of ELT. While AI tools such as ChatGPT have demonstrated considerable potential in automating assessment design, facilitating efficiency, and introducing novel test formats, concerns remain regarding the quality of output, particularly in distractor plausibility, construct validity, and linguistic nuance. In light of the above-mentioned studies, the following section will introduce the current extant AI tools that ELT practitioners could use for test item generation.

III. AI-BASED TOOLS FOR TEST ITEM GENERATION

This section will introduce four AI-based tools for automated item generation for L2 teaching. The first tool, ChatGPT, was initially trained on demonstration data to function as an AI assistant, followed by a stage of human-in-the-loop reinforcement where human raters evaluated the quality of its outputs. This process, known as reinforcement learning from human feedback (RLHF), enabled the model to generate responses approximating human-like communication. The current iteration, ChatGPT Plus, is powered by GPT-4-a

multimodal architecture capable of both interpreting and producing textual and visual content. While ChatGPT Plus requires a subscription, the freely available version based on GPT-3.5 remains accessible for purposes such as automated test item generation. As outlined in Shin (2023), generating test items with ChatGPT typically involves selecting a model item from an existing test and entering it into the prompt, often alongside instructions (e.g., "Create a new passage on a similar topic and generate corresponding multiple-choice items"). ChatGPT then produces an item modeled on the original, which the user can evaluate in terms of linguistic naturalness, coherence, and overall item quality. If the outcome is unsatisfactory, the prompt can be revised for further iteration. The following Figure 1 demonstrates the example of a prompt and output from ChatGPT.

FIGURE 1
Example of Automated Item Generation from ChatGPT-4



The second tool, Perplexity AI (Perplexity, 2025), shares functional similarities with ChatGPT in item generation but operates as an AI-driven conversational search engine. Utilizing OpenAI's GPT-4 and natural language processing techniques, Perplexity AI delivers responses grounded in real-time web content. One of its distinguishing features is the provision of source links, allowing users to verify information and access supporting materials. By offering access to multiple Large Language Models (e.g., GPT-4, Claude, and Mistral) and combining them with real-time search, Perplexity AI supports concise, source-based, contextually relevant, and verifiable responses (Garg, 2023). However, Perplexity AI did not provide source links or disclose information about its use of automated item generation. Perplexity AI generated a similar test item like Chat GPT when offering the same prompt as demonstrated in Figure 2.

FIGURE 2
Example of Automated Item Generation from Perplexity AI

Prompt:

Using the passage below as a model, create a new reading passage on a similar topic (e.g., environmental protection) targeted at high-intermediate EFL college students. Then, generate three multiple-choice comprehension questions that assess the main idea, specific details, and inference. Indicate the correct answer for each question

Output:

New Passage:

Planting trees is one of the simplest and most effective ways to combat environmental issues. Trees absorb carbon dioxide, a major greenhouse gas, and release oxygen, which is essential for life. Additionally, they provide habitats for countless species and help prevent soil erosion. Despite these benefits, deforestation continues at an alarming rate due to agriculture, urban development, and logging. To address this issue, many countries are launching reforestation projects and encouraging individuals to plant trees in their communities. By taking action now, we can ensure a healthier planet for future generations.

Comprehension Questions:

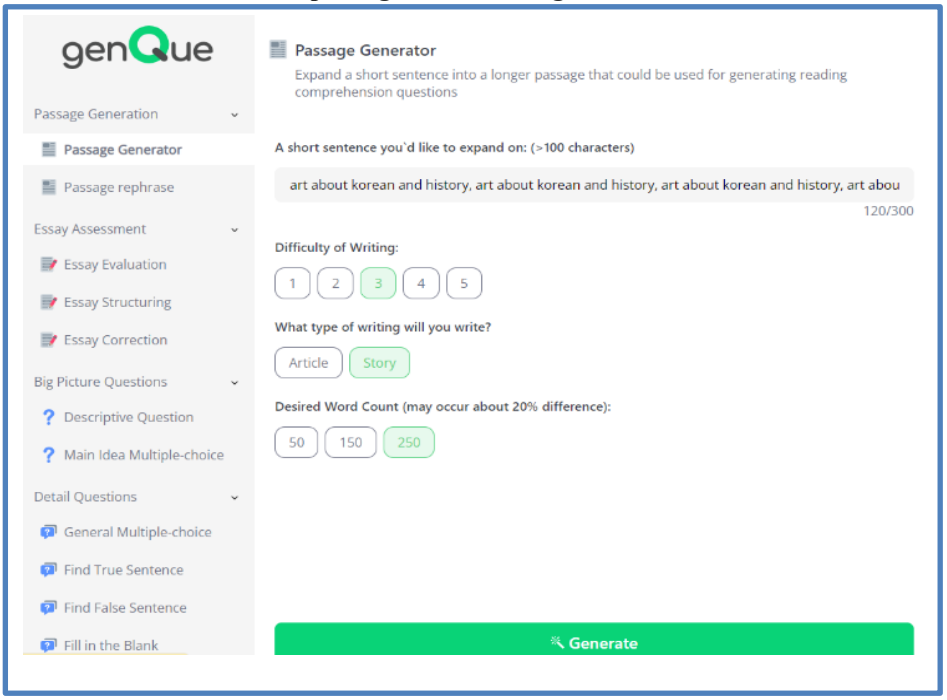
1. What is the main idea of the passage?

a) Urban development is the leading cause of deforestation.
b) Planting trees is an effective way to address environmental problems.
c) Soil erosion is caused by deforestation.
d) Reforestation projects are only conducted by governments.

Correct Answer: b) Planting trees is an effective way to address environmental problems.

The third tool, genQue, is an educational content generation platform that combines ChatGPT with proprietary technology developed by Artificial Society (2023). Unlike ChatGPT and Perplexity AI, genQue is not a search engine; rather, it is purpose-built to support the creation of English items for the CSAT. genQue offers a range of tools, including a passage generator and templates aligned with 12 distinct CSAT item types. When using the Passage Generator, users are required to input at least 100 characters as a passage seed, which may present usability challenges. The tool allows customization of difficulty level based on readability metrics and vocabulary complexity, as well as content type (fiction vs. nonfiction) and length (50, 150, or 250 words). Once the passage is generated, the item generator function can be used to produce test items across five differentiated levels of difficulty. Despite being a subscription-based platform, genQue offers significant utility for in-service elementary and secondary school teachers in Korea. It supports the generation of customized English reading passages and test items aligned with the level of the CSAT, features an editor conducive to educational applications, and facilitates the efficient production and management of instructional materials. Figure 3 illustrates a genQue's passage generator.

FIGURE 3
Example of genQue's Passage Generator

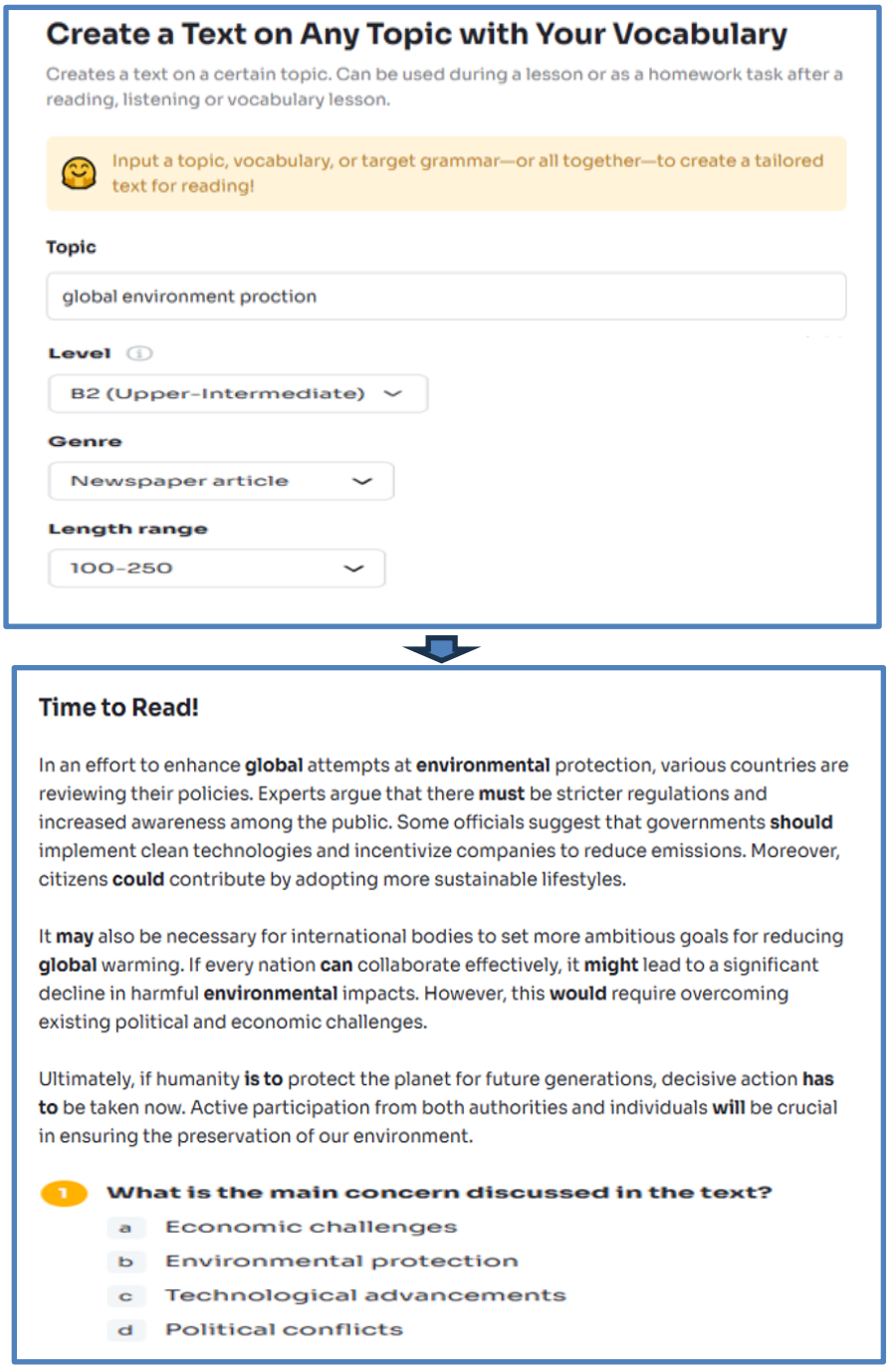


The fourth tool, Twee is an AI-assisted tool designed to facilitate automated item

generation by leveraging natural language processing and machine learning algorithms to produce pedagogically sound language learning materials (AI Tools for Teachers, Online Teaching Tools for English Educators, Twee, 2025). Developed with a focus on second and foreign language instruction, AI Twee allows educators and researchers to generate various types of language tasks, such as cloze tests, multiple-choice questions, and vocabulary exercises, with minimal manual input. Its user-friendly interface enables even non-technical users to customize item parameters based on proficiency levels, target linguistic features, and desired test formats. The tool draws from learner corpus data and general English usage corpora to enhance linguistic authenticity and cognitive validity, making it an increasingly valuable asset in computer-assisted language assessment.

In terms of automated item generation, AI Twee offers tailored item generation functions that align with their developmental stage in language acquisition. One particularly effective feature is its ability to generate context-rich cloze tests that focus on reading comprehension items based on instructor-provided texts. Upon inputting an expository or narrative passage, the tool can automatically generate multiple-choice comprehension questions, vocabulary-in-context items, and inference-based prompts. For instance, given a short reading text climate change, AI Twee may produce an item such as: "What is the main cause of recent global temperature increases, according to the passage?" *with options like* A) Natural climate cycles, B) Deforestation, C) Greenhouse gas emissions, D) Solar activity. The tool identifies key information units and aligns them with common question types found in standardized language assessments, such as main idea, detail, vocabulary, and inference questions. This kind of automated item generation not only saves time but also ensures consistency and appropriateness across test items. Moreover, the tool's capacity to adjust cognitive difficulty—by modifying sentence complexity or distractor plausibility—makes it especially useful for developing items that reflect the nuanced abilities of different L2 proficiency learners. This function provides instructors with pedagogically valid and cognitively appropriate test items, facilitating efficient and consistent assessment practices across instructional contexts. Figure 4 shows automated reading text and item generation according to L2 learners' language proficiency.

FIGURE 4
Example of Automated Text & Item Generation by Twee



Comparing each tool, ChatGPT-4 demonstrates remarkable flexibility in generating diverse L2 test items—ranging from lexical cloze items to higher-order analytical questions—thanks to its deep generative capacity and promptability. Research consistently highlights its potential: meta-analyses report substantial gains in learning outcomes ($g = 0.867$ for performance; $g = 0.457$ for higher-order thinking) when tasks are scaffolded using Bloom's taxonomy frameworks (Wang & Fan, 2025). However, significant limitations exist. The model is prone to hallucinations—fabricating details or referencing nonexistent sources—which undermines content validity (Jang et al., 2024). It may also perpetuate biases learned from training data, such as language or cultural biases, raising fairness and representativeness concerns. For classroom implementation, practitioners must rigorously vet each generated item for factual accuracy, bias, and psychometric alignment. Pedagogically, educators are encouraged to pair ChatGPT-generated prompts with well-defined instructional objectives and human-in-the-loop feedback mechanisms. This combination fosters authentic language use while maintaining educational rigor, mitigating the model's tendency to "oversimplify or misrepresent critical details" without oversight (Maity et al., 2024).

While ChatGPT exemplifies the flexibility and generative power of large language models, other AI-based tools such as Perplexity AI, genQue, and Twee offer alternative affordances tailored to different educational priorities and user needs. Perplexity AI excels at generating factual, citation-linked content quickly due to its integration of GPT-4 with real-time web search. Educators appreciate its traceable responses, which facilitate verification and factual check items (Garg, 2023). Yet the heavy reliance on surface-level retrieval limits its ability to produce pedagogically sequenced or cognitively demanding item sets; teachers cannot easily generate inference or multi-stage tasks in a single workflow. genQue and Twee offer stronger scaffolding via templates tuned to the CSAT item types or comprehension exercises.

Similarly, Twee leverages learner corpora and general English language data to generate pedagogically sound language tasks—including cloze tests and vocabulary exercises—tailored to learners' proficiency levels (Laura & Encarnación, 2025). Its user-friendly interface enables non-expert educators to produce context-rich reading comprehension items that reflect L2 learners' developmental stages. While Twee excels in generating cognitively appropriate and consistent items, it remains limited in flexibility compared to more open-ended generative models. From an instructional standpoint, teacher awareness of each tool's affordances is essential: Perplexity AI for rapid, verifiable fact-checking; genQue and Twee for standardized, readability-adjusted practice items. In all cases, educators should employ these tools within pedagogical frameworks—like adaptive Bloom's scaffolding—and integrate psychometric review to ensure fairness, mitigate bias, and enhance student interpretive depth.

Overall, ChatGPT and Perplexity AI offer notable advantages in terms of adaptability and

user control, as their open-prompt architecture allows for highly customized item generation tailored to specific learner profiles and instructional objectives. However, such flexibility presupposes a certain level of prompt engineering skill and pedagogical awareness on the part of the user, which may pose challenges for non-expert educators. In contrast, genQue and Twee provide more structured, template-driven environments that simplify the item creation process and reduce the cognitive load associated with task design. While this interface-based approach ensures consistency and accessibility, it may also constrain creativity and limit responsiveness to diverse learner needs. Thus, each platform reflects a trade-off between user autonomy and system scaffolding, underscoring the importance of aligning tool selection with the assessment context and user expertise.

IV. EVIDENCE-BASED APPROACHES FOR L2 EDUCATORS

Drawing upon the distinctive operational characteristics of the four AI-powered tools examined in the preceding section, this section presents pedagogically informed recommendations for language educators. genQue provides a comprehensive item generation framework that enables precise calibration of both passage complexity and question difficulty across five discrete proficiency levels. This granular control facilitates the development of assessment items that are systematically aligned with learners' developmental stages and specific pedagogical objectives.

A particularly noteworthy feature of genQue is its capacity to generate automated explanations for correct responses while simultaneously providing quantitative assessments of distractor effectiveness. This analytical capability enhances the tool's evaluative utility by enabling educators to systematically examine and refine the plausibility of alternative response options. Through this distractor analysis function, educators can optimize item quality by ensuring that incorrect options serve as meaningful cognitive challenges rather than trivial alternatives, thereby strengthening the overall diagnostic value of the assessment instrument. Given the distinct operational characteristics of each of the four AI-powered tools discussed in the preceding section, the following section seeks to provide pedagogically grounded recommendations for language educators. First, genQue offers a structured item generation system that enables users to fine-tune both passage and question difficulty across five discrete levels. This functionality facilitates the creation of test items that align with learners' proficiency levels and pedagogical needs. Notably, genQue provides automated explanations for correct answers and estimates the distractibility of distractor options. This distractor analysis enhances the evaluative potential of the tool by allowing users to gauge and refine the plausibility of alternative responses, thereby contributing to improved item quality.

Despite these strengths, genQue is limited by its design parameters. Specifically, the platform is currently restricted to generating English reading comprehension items aligned with the Korean CSAT framework. It does not support the creation of alternative item types such as listening or speaking tasks. Moreover, the passage length is constrained to approximately 250 words, and the content difficulty is capped at the CEFR B2 level (high-intermediate)[1], which restricts its applicability for advanced learners (C1–C2). Nevertheless, genQue's fine-grained level control arguably surpasses that of ChatGPT or PerplexityAI, where difficulty calibration relies on broader prompt modifiers such as "more difficult" or "CEFR B2 level." While this allows for general adjustment, it lacks the precision of genQue's internally tiered levels.

Compared to genQue, which is tightly aligned with a specific test format (i.e., the Korean CSAT reading section) and offers limited item types, AI Twee is particularly advantageous for L2 practitioners seeking a flexible, classroom-oriented tool that supports a variety of item types, including cloze tasks, grammar, vocabulary-in-context items, and inference questions. Unlike genQue, Twee not only allows instructors to upload their own reading passages, but also creates reading passages from a variety of genres, such as fiction, descriptive, newspaper, and argumentative articles, and instantly generates contextually grounded comprehension questions according to cognitive difficulty. This makes Twee especially useful for task adaptation and formative assessment in EFL classrooms, where materials often need to be tailored to diverse proficiency levels and topical relevance. Furthermore, Twee's ability to subtly adjust cognitive difficulty—such as manipulating distractor plausibility or syntactic complexity—enables the creation of pedagogically valid items that align with learners' developmental stages in language acquisition.

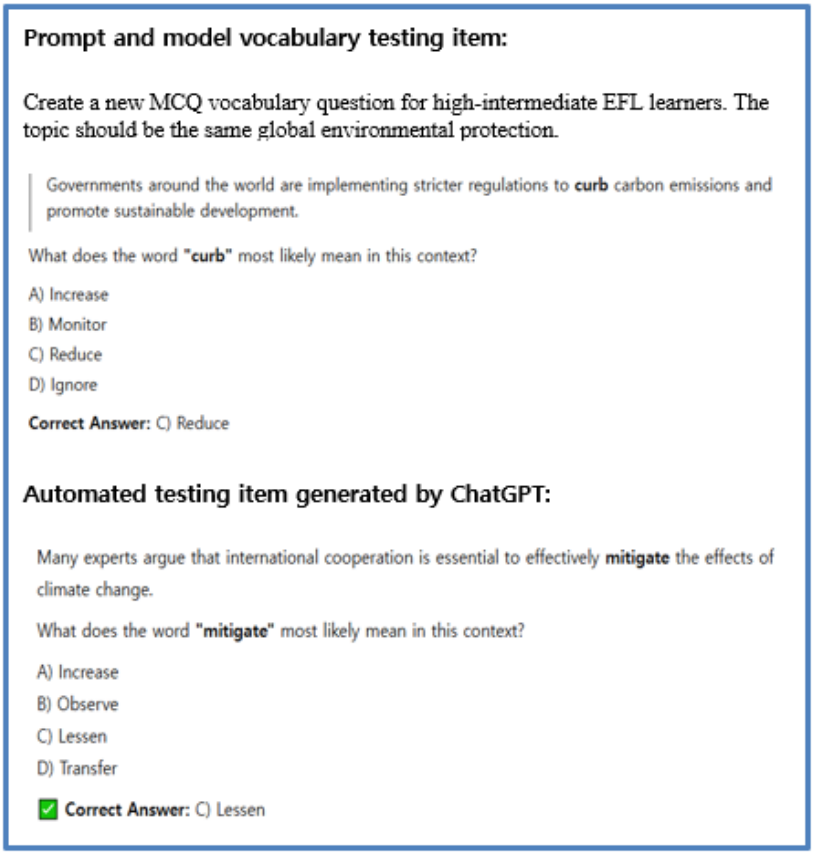
By contrast, genQue's strength lies in its alignment with high-stakes test design. Its item generator is built around the standardized formats and difficulty gradations of the CSAT, offering templates for 12 distinct item types and passage generation functions based on readability and vocabulary metrics. While this ensures a high degree of control and reliability for summative assessment contexts, it also limits the scope of item types and user autonomy. For example, inputting a passage seed of at least 100 characters may present usability challenges for instructors unfamiliar with prompt engineering. Additionally, genQue's coverage is restricted to reading items within CEFR B2-level complexity, whereas Twee

[1] The Common European Framework of Reference for Languages (CEFR), developed by the Council of Europe, provides comprehensive language proficiency across a range of skills, including listening, speaking, reading, and writing. The CEFR comprises six levels—A1, A2, B1, B2, C1, and C2—which represent a progression from basic user (A1, A2) to proficient user (C1, C2). This scale serves as a standardized benchmark for language teaching, assessment, and curriculum design, enabling cross-national comparability in educational and professional settings (Council of Europe, 2001).

offers a broader range of item types, albeit with less control over formal test specifications. Therefore, L2 educators engaged in curriculum-integrated assessment or differentiated instruction may find Twee more accessible and adaptable for everyday teaching purposes, whereas genQue may be better suited for test-aligned material development in exam-focused settings.

Conversely, ChatGPT and PerplexityAI offer greater flexibility in terms of test item variety. Through prompt-based interaction, these tools can generate a wide range of item types—including vocabulary, pronunciation, listening, and integrated tasks—depending on the user's instructional goals and language focus. The following Figure 5 presents an example of a vocabulary test item generated by ChatGPT to illustrate this affordance.

FIGURE 5
Prompt & Example of Vocabulary Testing Item



When the initial prompt was entered into Perplexity AI, it produced questions and

responses that were highly similar to those generated by ChatGPT. Notably, the contextual framing and the format of the answer choices closely resembled those of ChatGPT, which was particularly striking. These findings suggest that Perplexity AI demonstrates comparable performance to ChatGPT in the context of AI-generated test item construction.

Another critical consideration lies in the way prompts are phrased during their formulation. For instance, the researcher's observations indicate that when the prompt included the word 'passage' (e.g., "Create a new passage with a different topic and make new multiple choices as follows"), ChatGPT failed to generate the intended reading passage and test items, despite successfully generating other types of test items. However, when the term 'passage' was replaced with 'text', ChatGPT produced the desired output correctly. A similar issue was observed with Perplexity AI, highlighting limitations in prompt interpretability. These inconsistencies suggest that users must be mindful of the variability and instability in how prompts are processed by both ChatGPT and Perplexity AI. In contrast, such concerns are less prominent when using genQue and Twee, as they employ a built-in question generation template that minimizes ambiguity in prompt formulation.

Nevertheless, ChatGPT and Perlexity AI illustrate the flexibility of chat-based prompting, whereby users can modify their inputs to elicit the desired output. In other words, if the AI does not fulfill a task as intended, users can refine the prompt to guide the model toward generating the targeted content.

TABLE 1
AI-Based Tools for Test Item Generation: Features, Limitations, and Pedagogical Suitability

Tool	Key Features	Strengths	Limitations	Best Suited For
genQue	- Structured item generation aligned with Korean CSAT - Difficulty calibration for passages and questions - Distractor analysis and explanation	- Fine-grained difficulty control - Diagnostic distractor evaluation - Templates for 12 CSAT item types	- Limited to reading items only - CEFR B2 ceiling- Fixed format focused on CSAT	- Teachers developing standardized test materials - Educators preparing CSAT-style reading tasks - Instructors focused on summative assessments
Twee	- Supports various item types: cloze, vocabulary, MCQs - User-supplied or generated texts - Genre and difficulty customization	- Broad item type support - Adjusts cognitive complexity and genre - Accessible for non-experts	- Less alignment with formal test specifications - Limited psychometric control	- L2 teachers designing classroom-based or formative assessments - Instructors needing genre-diverse, adaptive language tasks

ChatGPT-4	- Open-ended prompt-based generation - Capable of producing diverse item types (listening, pronunciation, integrated tasks)	- Highly flexible and creative-Supports a wide range of formats - Strong natural language generation	- Requires prompt engineering skill-Risk of hallucinations or vague outputs - No built-in pedagogical scaffolding	- Tech-savvy educators familiar with AI prompting - Creative item designers in exploratory or task-based contexts
Perplexity AI	- Combines GPT-4 with real-time web search-Provides cited responses with source links - Chat-based search interface	- Quick, verifiable responses-Supports factual Q&A generation - Comparable to ChatGPT in language item generation	- Limited cognitive depth - Not optimized for pedagogical task design - Sensitive to prompt wording	- Educators seeking reliable, fact-based content - Instructors needing source-backed support items - Quick information-based question generation use cases

V. CONCLUSION

This study presented four distinct AI-powered tools for automated item generation, outlining their key functionalities and offering pedagogical recommendations for ELT professionals. Consistent with previous research (Kurdi et al., 2020), the researcher affirms that AI-assisted automatic item generation holds considerable promise for test developers, including classroom teachers tasked with creating assessment materials. Furthermore, AI-based tools for automatic test generation are anticipated to become integral to large-scale and continuous testing practices (Circi et al., 2023), as they enable the development of extensive item banks in a cost-effective and time-efficient manner, with minimal human intervention.

Despite these advantages, the current capabilities of AI necessitate continued involvement from human experts to enhance the quality and validity of AI-generated test items. Nevertheless, the researcher argues that language educators will increasingly require training in AI-assisted item development, particularly with tools such as those introduced in this study. In addition, future research should investigate how AI-generated test items are received by learners, teachers, and other stakeholders. Finally, empirical studies assessing the practical utility of automated item generation in diverse educational contexts are warranted. With these ongoing efforts, the researcher contends that automated test generation has the potential to significantly transform language testing and assessment.

REFERENCES

AI Tools for Teachers, Online Teaching Tools for English Educators, Twee. (2025). *Twee.* http://app.twee.com/my-lessons/70e1e6b0-f811-45b9-acda-c481884a38a8

Artificial Society. (2023). *genQue*. https://www.genque.ai/

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 903077. https://doi.org/10.3389/frai.2022.903077

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers & Education: Artificial Intelligence, 5*, 100161.https://doi.org/10.1016/j.caeai.2023.100161

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

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, 858273. https://doi.org/10.3389/feduc.2023.858273

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Garg, S. (2023). What is Perplexity AI and how to use it. *All Things How*. https://allthings.how/what-is-perplexity-ai-and-how-to-use-it/

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726-733.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757-765.

Ha, D. N. L., & Nguyen, A. T. (2024). Artificial intelligence-based assessment in ELT exam creation: A case study of Van Lang University lecturers. *Saudi Journal of Language Studes, 5*(1), 34-41.

Jang, H. Y., Lee, J., & Cho, Y. (2024). Hallucination in large language models and its implications for education. *AI & Education Review, 4*(1), 33-50.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537-550.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*, 121-204.

Laura, P. Gómez., & Encarnación, P. Pinazo. (2025). Artificial intelligence in the training of public service interpreters. *Language & Communication*, *103*, 86-107.

Luo, Y., Wei, W., & Zheng, Y. (2022). Artificial intelligence-generated and human expert-designed vocabulary tests: Comparative study. *SAGE Open, Januaray-March*, 1-12.

Perplexity AI. (2025). *Perplexity AI*. http://www.perplexity.ai/

Maity, S., Deroy, A., & Sarkar, S. (2024). Evaluating factual consistency and semantic coherence in LLM-generated educational content. *Educational Technology Research and Development, 72*(2), 225-243.

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high-quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning, 15*, 1-13.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1-24.

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.

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.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847-857.

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Meta-analysis. *Humanities and Social Sciences Communications, 12*, Article 621.

Examples in: English
Applicable Language: English
Applicable Levels: Tertiary

Youn-Kyoung Lee (researcher)
Department of English Language and Literature
Kyungpook National University
41566 Daehakro 88, Bukgu, Daegu, Republic of Korea
Email: yklee0805@knu.ac.kr

Received : July 7, 2025
Reviewed : September 2, 2025
Revised version received : September 23, 2025