Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

Development and Validation of a Human-in-the-Loop (HITL) Model for AI-Based Automatic Item Generation in the CSAT Korean Reading Section

Suji Lee
Korea National University of Education

*Corresponding Author: Suji Lee, sasa0802@naver.com

ABSTRACT

This study develops and validates a Human-in-the-Loop (HITL) model for generative AI-based Automatic Item Generation (AIG) of College Scholastic Ability Test (CSAT) Korean reading items. As a high-stakes assessment of higher-order thinking, the CSAT demands rigorous reliability and validity (AERA, APA, & NCME, 2014; Kwon et al., 2017), yet hand-crafted development is resource-intensive and over-relies on developers' assessment literacy (Jung, 2008; Rudner, 2009). Generative AI can ease these constraints but, used alone, is limited by hallucination and weak distractors, requiring expert supervision (Haladyna et al., 2002; Hommel et al., 2022). The HITL model therefore integrates AI efficiency with the educational and psychometric judgment of human experts (Burke, 2025; Sayin & Gierl, 2024). A 10-step protocol integrated CSAT development principles, the 2015 revised Korean curriculum, AIG theory, and Chain-of-Thought prompting (Wei et al., 2022), pairing an expert-alignment phase with a backward-design phase for creating, reviewing, and refining passages and items. Applying it, one researcher generated three passage-item sets (12 items in total) across three domains in 24 hours. Validation drew on expert ratings from 18 in-service Korean teachers and responses from 401 twelfth-grade students, examining readability, inter-rater reliability, CTT and IRT properties, and congruence between expert judgment and empirical performance. Generated passages showed no significant readability difference from past CSAT passages and scored high for formal completeness ($M > 4.8$) but lower for difficulty calibration and assessment suitability ($p < .05$). Items showed acceptable properties ($\alpha = .656$, mean $a = 1.097$) but limited discrimination for critical-reading items (mean a = 0.504) and weak distractor attractiveness. Notably, the passage rated highest by experts for educational quality yielded the lowest mean item discrimination ($\rho = -1.0$). These findings show that AIG improves structural completeness, but human intervention and data-driven validation remain indispensable for psychometric rigor and educational validity, redefining the assessment expert as a psychometric designer and data-driven coordinator.

Keywords: Automatic Item Generation (AIG), Passage Generation, Human-in-the-Loop (HITL), College Scholastic Ability Test (CSAT), Reading Domain, Reading Assessment, Classical Test Theory (CTT), Item Response Theory (IRT)

Check for updates

OPEN ACCESS

Brain, Digital, & Learning
2026, Vol. 16, No. 2, 165–187
https://doi.org/10.31216/BDL.2026.16.2.4

Received: June 19, 2026
Revised: June 30, 2026
Accepted: June 30, 2026

Introduction

This study focuses on the reading domain of the College Scholastic Ability Test (CSAT) Korean language area, a high-stakes standardized assessment that profoundly impacts the entire educational ecosystem by measuring higher-order thinking skills (American Educational Research Association [AERA] et al., 2014; Kwon et al., 2017). Reading proficiency is a foundational competency that sustains general learning, functioning as a mechanism for "reading to learn"—acquiring new information and expanding thought (Chall, 1983; Korean Reading Association, 2021). Consequently, CSAT reading items serve as a core indicator for evaluating cross-curricular thinking and information processing capabilities (Kim, 2015). However, the traditional hand-crafted item development process exhibits structural limitations; it imposes substantial time and financial burdens (Rudner, 2009) and over-relies on individual developers' assessment literacy, making it difficult to ensure consistent item quality (Jung, 2008; Luecht, 2012).

Generative AI-based Automatic Item Generation (AIG) has emerged as a promising alternative to address these limitations. The advent of large language models (LLMs) has significantly lowered the technical barriers to item generation by enabling operation through natural language prompts alone (Burke, 2025). Furthermore, the CSAT provides an ideal ground truth for AI training, given its accumulation of high-quality item data over 30 years. Nevertheless, utilizing AI as a standalone tool presents inherent limitations, including hallucination phenomena, logical leaps, and critically, poor distractor attractiveness (Haladyna et al., 2002; Hommel et al., 2022). Consequently, the Human-in-the-Loop (HITL) model has surfaced as a practical alternative, cyclically integrating AI's generative efficiency with the educational and psychometric judgment of human experts throughout the item development process (Burke, 2025; Sayin & Gierl, 2024).

A synthesis of prior research indicates that AIG has evolved from template-based approaches to generative models (Brown et al., 2020; Gierl & Lai, 2012; Song et al., 2025), with generation quality improving significantly through prompt engineering strategies such as Chain-of-Thought (CoT) and few-shot prompting (Scaria et al., 2024; Wei et al., 2022). Although some studies have demonstrated the feasibility of integrated passage and item generation (Attali et al., 2022; Bezirhan & von Davier, 2023), distinct limitations persist in securing distractor attractiveness and measuring higher-order thinking skills (Lin & Chen, 2024; Shin & Lee, 2023; Wen & Chu, 2025). Most importantly, the lack of rigorous psychometric validation using large-scale learner data remains a critical research gap (Gorgun & Bulut, 2024; Tan et al., 2025).

Addressing this gap, this study aims to empirically validate generative AI-based item development by designing an HITL protocol specialized for the CSAT Korean reading domain. The quality of the generated passages and items is evaluated by cross-validating large-scale student response data (N = 401) and expert ratings (N = 18) through Classical Test Theory (CTT) and Item Response Theory (IRT) analyses. Specifically, this study seeks to overcome the methodological limitations of previous research through integrated passage and item generation, the application of systematic collaborative procedures, and empirical data-driven validation.

To achieve these objectives, the following research questions were established:

1. How should an HITL protocol be designed and implemented to reflect the assessment requirements and curriculum achievement standards of the CSAT Korean reading domain?

2. What is the quality of the passages and items generated through the developed HITL protocol?

3. What implications do the comprehensive results of expert ratings and student response data analysis provide regarding the effectiveness of the HITL model and the role of human experts?

Through these research questions, this study intends to empirically establish the validity of generative AI-based item development and provide theoretical and practical grounds for redefining the role of assessment experts in the AI era.

Methods

Research Procedure

This study aims to develop an HITL protocol that automatically generates passages and items for the CSAT Korean reading domain using a Human-in-the-Loop (HITL) model, and to empirically validate the quality of the outputs. To this end, the research was conducted following a systematic five-step procedure: Theoretical Foundation Exploration → Protocol Design → Protocol Implementation → Quality Validation → Comprehensive Analysis and Interpretation.

In the Protocol Design phase, the theoretical foundations of CSAT reading assessment requirements, AIG principles, and the HITL model were established. These were structured into actionable, step-by-step collaborative procedures and prompt systems. Additionally, an analytical rubric comprising seven criteria each for passages and items was developed to serve as both a review standard and a quality validation tool.

In the Protocol Implementation and Outcome Generation phase, a single researcher utilized GPT-4 and Claude 3 Opus to generate passages and items using a backward design approach. A multi-layered refinement process was conducted, involving AI cross-review followed by a final review by a human expert. Consequently, passage-item sets (totaling 12 items) across three domains—humanities/arts, social sciences/culture, and science/technology—were produced within a 24-hour timeframe.

In the Quality Validation phase, a dual approach of quantitative analysis and expert rating was applied to the generated passages and items. Passages were evaluated using expert ratings and the KReaD index to measure text readability. Items were analyzed using CTT and IRT based on actual response data from 401 high school seniors, which was subsequently cross-validated with the expert ratings.

Finally, in the Comprehensive Analysis and Interpretation phase, the expert ratings and student response data were integrated to provide an in-depth diagnosis of the technical strengths and limitations of generative AI. Based on these findings, the role of human experts was defined, and strategies for advancing the protocol were derived.

The overall procedure of this study is outlined below.

Fig. 1. Research procedure of HITL-based AIG for CSAT reading items

Participants

Participants in this study consisted of the protocol implementer, an expert rating panel, and a student response group. To verify practical applicability under realistic conditions, protocol implementation was conducted by a single in-service Korean language teacher, who possesses eight years of teaching experience at academic high schools. The expert rating panel, organized to evaluate the quality of the generated passages and items, comprised 18 in-service middle and high school Korean language teachers with high school teaching experience. Their demographic distribution is presented in Table 1. Student response data were collected from 12th-grade students across four academic high schools located in Chungcheongbuk-do. From an initial pool of 440 datasets, 39 incomplete or unengaged responses were excluded, yielding a final sample of 401 datasets for analysis. The sampling details are outlined in Table 2.

Table 1. Distribution of expert raters (*N* = 18)

School Level	Middle School		High School		Total
Gender / Teaching Experience	Male	Female	Male	Female	
1–5 years	-	1	1	3	5
6–10 years	-	5	2	3	10
11+ years	1	-	1	1	3
Total	1	6	4	7	18

Table 2. Distribution of student participants (*N* = 401)

Category	Sub-category	*N*	%
Gender	Female	243	60.6
	Male	158	39.4
Total		401	100
Academic Level (Teacher Rating)	High	102	25.4
	Medium	170	42.4
	Low	129	32.2
Total		401	100

Instruments

The HITL protocol, serving as the core instrument of this study, consists of procedural steps and stage-specific prompts through which a human expert systematically intervenes in the AI generation process to enhance item quality. The protocol is divided into a design phase (①~⑤) for establishing assessment criteria and a generation phase (⑥~⑩) for producing and refining the outputs.

In the design phase, the AI is aligned as an assessment expert through persona configuration, assessment context setting, curriculum and test guideline learning, and past exam set learning. This ensures the logical coherence and curriculum alignment of the generated outputs. Past exam sets were limited to the most recent five years, as CSAT testing trends shift slightly each year and recent items best reflect current conventions. Moreover, because Step ⑤ relies on prompt-based exemplar learning rather than large-scale pretraining, supplying too many examples increases processing burden. The five-year scope was therefore set to balance recency and representativeness while keeping the prompt load manageable. In the generation phase, passages are generated based on a backward design approach. Subsequently, passages and items are finalized through multi-layered refinement, which iteratively applies AI cross-review (GPT ↔ Claude) and the human expert's final judgment. The overall structure is presented in Table 3.

Table 3. HITL-based passage and item generation protocol

Protocol Phase		Prompt Content
1. Design	① Persona Setting	● Define role: CSAT Korean reading item developer and quality control expert ● Establish core expertise: 15 years of teaching experience, reading education/assessment research ● Limit performance scope: Passage creation by domain, item creation by cognitive type
	② Assessment Context Setting	● Establish basic assessment conditions: Target, purpose, format, scale ● Set passage development conditions: Domain, length, difficulty, rationale ● Set item development conditions: Structure, difficulty, cognitive level, rationale
	③ Curriculum Learning	● Input 2015 revised reading curriculum achievement standards, learning elements, and testing points (reading methods and domains)
	④ Test Guideline Learning	● General assessment guidelines: Basic principles, cognitive type definitions, domains ● Passage creation guidelines: Length, topic selection, exclusion criteria ● Item creation guidelines: Stem, distractors, <Context> drafting principles ● Review criteria learning
	⑤ Past Exam Set Learning	● Input 5-year past CSAT Korean reading passage and item sets ● Instruct retrieval based on analytical criteria (passage structure, item logic, distractor construction principles)
2. Generation	⑥ Passage Generation	● Select/confirm topic domain and pre-conceive item elements ● [Backward Design] Create passage optimized for item elements ● Learn and adhere to passage rubric ● GPT self-review and revision
	⑦ Passage Review and Revision	AI Review ● [Claude] Learn rubric & review passage → [GPT] Revise based on review Teacher Review ●[Teacher] Review based on rubric → [GPT] Revise based on review
	⑧ Item Generation	● Confirm and output final passage ● Generate 4 items based on the final passage ● Recall test guidelines ● Learn item rubric ● GPT self-review and revision
	⑨ Item Review and Revision	AI Review ●[Claude] Learn rubric & review items → [GPT] Revise based on review Teacher Review ●[Teacher] Review based on rubric → [GPT] Interactive, item-by-item revision with teacher
	⑩ Final Set Output	● Output final passage-item set and explanatory materials

The analytical rubrics utilized in this study are multidimensional tools for evaluating passage and item quality, designed to rate seven criteria each on a 5-point Likert scale. Drafts of the rubrics were formulated based on prior research analyzing the psychometric quality factors of Korean language assessment items (Kwon et al., 2017; Nam et al., 2022; Chung et al., 2022). These drafts were finalized following a review by one professor of Korean language education and three in-service teachers with over seven years of experience. These rubrics were consistently utilized not only as review standards for the AI and the human expert during the generation process but also as the expert rating tools for the final outputs. The passage quality rubric is presented in Table 4, and the item quality rubric is presented in Table 5.

Table 4. Analytic rubric for passage quality

Evaluation Target	Evaluation Criterion	Description	Rating Scale
Passage	1. Appropriateness of Difficulty	Are the vocabulary level, sentence complexity, and required background knowledge appropriate for 12th-grade learners?	1-5
	2. Accuracy of Content	Is the presented information consistent with general knowledge, non-contradictory, and coherent?	1-5
	3. Unity of Theme	Does the text consistently address a single central theme throughout, without extraneous information?	1-5
	4. Systematicity of Structure	Is the overall structure logical, and are the connections between sentences and paragraphs natural and smooth?	1-5
	5. Clarity of Explanation	Are the explanations of core concepts clear, and are specific examples or supplementary explanations appropriate?	1-5
	6. Assessment Suitability	Is it possible to develop a diverse range of items measuring literal, inferential, critical, and creative comprehension?	1-5
	7. Accuracy of Expression	Are spelling, grammar, and vocabulary selection accurate, without awkward expressions or errors?	1-5

Table 5. Analytic rubric for item quality

Evaluation Target	Evaluation Criterion	Description	Rating Scale
Item	1. Curriculum Alignment	Does the item validly measure the targeted curriculum content elements (literal, inferential, critical, and creative reading)?	1-5
	2. Explicitness of Assessment Elements	Are the targeted assessment elements clearly stated so that learners can accurately understand the task?	1-5
	3. Promotion of Thinking Skills	Does the item require cognitive skills appropriate for the targeted content, promoting the thinking process rather than simple memorization?	1-5
	4. Appropriateness of Difficulty and Discrimination	Is the difficulty suitable for 12th-grade learners, and does it effectively discriminate between high- and low-achieving students?	1-5
	5. Clarity and Logicality of Information Presentation	Is the information in the stem, <Context>, and options presented clearly without confusion, and is it logically systematized?	1-5
	6. Attractiveness of Distractors	Are the distractors attractive to learners without overlap or interference among the options?	1-5
	7. Formal Completeness	Are all elements grammatically, lexically, and structurally accurate, and is formal balance and consistency maintained among the options?	1-5

The student response test utilized the three passage-item sets (totaling 12 items) finalized through the HITL protocol. Each set consisted of one informational passage linked to four five-option multiple-choice items, specifically designed to measure literal, inferential, critical, and creative reading abilities based on the reading achievement standards of the 2015 revised curriculum. Notably, items assessing critical and creative reading included a <Context> block to present alternative perspectives or specific case applications. The test duration was set to 40 minutes, scaled to the three passages and twelve items in proportion to the average time generally required to solve a single reading set (one passage with three to four items, roughly 10–13 minutes) in the actual CSAT Korean section. This test instrument provided the empirical data for analyzing psychometric properties such as item difficulty and discrimination. The structural composition of each set is detailed in Table 6.

Table 6. Composition of the passage–item sets used for the student response test

Set	Topic Domain	Passage Theme	Items	Assessment Content (Cognitive Level)
1	Huma nities & Arts	Kant's and Ar istotle's criteria for moral judgment	4	① Verifying factual detail (Literal) ② Inferring implications of core concepts (Inferential) ③ Critiquing from an alternative perspective in <Context> (Critical) ④ Applying to a specific case in <Context> (Creative)
2	Social Sc iences & C ulture	Platform labor and issues within the social security system	4	① Identifying specific information (Literal) ② Inferring from presented information (Inferential) ③ Critiquing from the perspective in <Context> (Critical) ④ Applying to a specific case in <Context> (Creative)
3	S cience & Tech nology	L imitations of generative AI and principles of the HITL model	4	① Understanding core principles (Literal) ② Inferring relationships between concepts (Inferential) ③ Evaluating from conflicting perspectives in <Context> (Critical) ④ Applying to a specific case in <Context> (Creative)

Data Analysis

Data analysis was conducted using R statistical software (Item Response Theory (IRT) via the mirt package; Classical Test Theory (CTT) and Intraclass Correlation Coefficient (ICC) via psych; Analysis of Variance (ANOVA) and correlation analysis via stats; and data visualization via ggplot2). Passage readability was measured using the KReaD index, a Korean readability measure that combines lexical and syntactic features—mean sentence length, sentence-structure score, complex-sentence ratio, vocabulary diversity, and difficult-vocabulary ratio—through a graded vocabulary list and an R-based program (Cho & Lee, 2020). Unlike English formulas such as Flesch, it was developed to reflect the linguistic properties of Korean, and was adopted here to compare the generated and past passages on a common scale. Mean differences between the generated passages and the past five years of CSAT passages were then evaluated using an independent samples t-test.

The reliability of the expert ratings was examined across two dimensions. The internal consistency of the rubrics was calculated using Cronbach's α and McDonald's ω, while inter-rater reliability was determined via ICC(2,k), applying the criterion of ≥ .75 as indicative of good reliability (Koo & Li, 2016). Between-group differences in rating scores were tested using one-way and two-way mixed-effects ANOVA, with significant findings subjected to Tukey's Honest Significant Difference (HSD) post-hoc test ($p < .05$).

The psychometric properties of the items were analyzed concurrently utilizing both CTT and IRT frameworks. Under CTT, item difficulty (proportion correct, target range: .20–.80), item discrimination (point-biserial correlation, threshold: ≥ .30), and test reliability (Cronbach's α, threshold: ≥ .70) were computed. Under IRT, the discrimination (*a*) and difficulty (*b*) parameters were estimated applying the Two-Parameter Logistic (2PL) model. The discrimination parameter was interpreted based on the criteria established by Baker (2001) and Seong, T (2019): 0.65–1.34 as 'moderate', 1.35–1.69 as 'high', ≥ 1.70 as 'very high', and < 0.65 as 'low'. Model fit was evaluated using M2 and the Root Mean Square Error of Approximation (RMSEA), while individual item fit was verified using S-χ^2. Measurement precision was assessed through the Test Information Function (TIF). Finally, Pearson and Spearman correlation analyses alongside multiple regression analysis were conducted to determine the congruence between expert ratings and empirical student responses.

Results and Discussion

Design and Implementation of the HITL Protocol

The HITL protocol developed in this study consists of a "Design Phase" (①~⑤), wherein a human expert establishes the learning environment for the AI, and a "Generation Phase" (⑥~⑩), wherein the expert proactively intervenes to produce and refine the outputs. Each phase features a distinct division of roles between the user (teacher) and the AI (GPT-4, Claude 3 Opus) while maintaining an interactive structure.

The core of the Design Phase (①~⑤) is "conditioning" the AI to function not merely as a text-generating tool, but as an assessment expert. The researcher explicitly established the criteria and constraints for generation by defining the persona and assessment context, and by inputting curriculum and test guidelines. Furthermore, by providing past 5-year CSAT data using a few-shot approach, the AI was induced to learn the implicit patterns of item development and option construction that are difficult to reduce to explicit rules. This constitutes a core design strategy to ensure the authenticity of the generated outputs.

The Generation Phase (⑥~⑩) is characterized by continuous human expert intervention and quality control. First, in the passage generation step, "backward design" was applied; the reading comprehension components to be measured were established in advance and reflected in the passage structure, thereby securing alignment between the passages and items. Subsequently, in the review step, formal errors were initially filtered out through GPT's self-review and Claude's cross-review. Following this, the human expert made the final judgments based on educational validity and psychometric quality. These dimensions were applied operationally through the criteria of the analytic rubric (Table 5): educational validity through curriculum alignment, explicitness of assessment elements, and promotion of thinking skills, and psychometric quality through the appropriateness of difficulty and discrimination and the attractiveness of distractors. The same rubric thus functioned consistently as both the review standard during generation and the final rating tool. Specifically, during the item generation step, the human expert directly adjusted these measurement elements to compensate for the AI's technical limitations.

This structure is significant as it is not a simple division of labor, but a systematic collaborative model where the AI's efficiency and the human expert's proficiency are cyclically integrated. Specifically, the AI maximizes productivity through draft generation and initial review, while the human expert ensures output validity through higher-order judgments. The roles and key characteristics of each protocol phase are detailed in Table 7.

Table 7. Stepwise roles and key features of the HITL protocol

Phase	Protocol	Role and Interaction Flow	Key Characteristics of the Protocol
1. Design	① Persona Setting	[Teacher] Define expert identity (experience, expertise) and limit role → [GPT] Learn expert role/expertise and establish identity	• Assessment Expert Role Construction: Conditioning the AI as an expert who has internalized the researcher-designed assessment principles.
	② Assessment Co ntext Setting	[Teacher] Define assessment target/purpose, passage/ item format requirements → [GPT] Internalize output s pecifications/ constraints	• S pecification-BasedDesign: Pre-determining output specifications and target quality levels.
	③ Curriculum Learning	[Teacher] Input 2015 revised reading achievement standards → [GPT] Learn achievement standards, learning elements, and assessment elements	• Alignment with Curriculum: Limiting the criteria scope to align with public education goals and content systems.
	④ Test Guideline Learning	[Teacher] Input KICE CSAT item development manual → [GPT] Internalize guidelines for passages, stems, options, and <Context>	• Internalization of High-Stakes Assessment Criteria: Ensuring fairness, validity, and reliability.
	⑤ Past Exam Set Learning	[Teacher] Input 5-year past CSAT reading sets → [GPT] Analyze and learn test trends, writing style, and logical structures	• Data-Driven Pattern Learning: Acquiring implicit patterns such as actual test trends and styles.
2. Generation	⑥ Passage Generation	[Teacher] Select topic domain and command generation → [GPT] Generate draft passage (V1.0) based on backward design	• Application of Backward Design: Structuring item assessment elements prior to passage generation.
	⑦ Passage Review and Revision	[GPT] Self-review (V1.0 → V1.1) → [Claude] Cross-review → [GPT] V1.2 → [Teacher] In-depth review → [GPT] Final version (V1.3)	• Multi-layered Evaluation (Passage): Combining AI's check of form/logic/facts with human's final judgment of content/educational validity.
	⑧ Item Generation	[Teacher] Command item generation based on final passage → [GPT] Generate draft of 4 items (V1.0)	• Strict Passage-Based Generation: Constructing correct/incorrect options relying solely on evidence within the passage.
	⑨ Item Review and Revision	[GPT] Self-review (V1.0 → V1.1) → [Claude] Cross-review → [GPT] V1.2 → [Teacher] Review item difficulty, discrimination, and distractors, then 'Pass' → [GPT] Final version (V1.3)	• Multi-layered Evaluation (Item) & Human-led Refinement: Human directly adjusts difficulty, discrimination, and distractor attractiveness.
	⑩ Final Set Output	[Teacher] Command final output generation → [GPT] Automatically output student assessment materials + teacher explanatory materials	• Automated Packaging: Generating immediately usable outputs without editing.

Furthermore, to verify protocol efficiency, a program automating steps ① through ⑥ was experimentally implemented. When the user inputs basic settings, the process from persona setting to draft passage generation runs automatically. In the subsequent review phase, it operates as an interactive structure where the human expert directly intervenes to decide on revisions and "pass" status. This design maintains a balance between automation and expert control, securing AIG's efficiency while enabling quality control.

Consequently, this protocol does not merely separate AI and human roles but specifically proposes an item generation system applicable to high-stakes assessments through an interactive integration across the entire process of generation, review, and refinement.

Item Generation and Refinement

This section analyzes the multi-layered refinement process of the final item sets produced based on the confirmed passages. Item generation proceeded through the following stages: GPT draft (V1.0) → GPT self-review (V1.1) → Claude cross-review (V1.2) → Human expert final version (V1.3). This incremental process functioned as a necessary mechanism to secure psychometric quality, going beyond mere structural completeness and logical validity.

Items in the initial stage (V1.0) acquired the superficial form of a four-level cognitive hierarchy (literal, inferential, critical, and creative reading); however, the attractiveness of the distractors—the core of discrimination—was significantly low. Distinguishing between correct and incorrect options relied on simple content contradiction with the passage or obvious factual errors, enabling correct answer identification without complex reasoning. This indicates that while AI is adept at reproducing the superficial format of items, it has inherent limitations in psychometric design intended to elicit sophisticated distractor responses.

The subsequent GPT self-review (V1.1) and Claude cross-review (V1.2) primarily supplemented logical consistency and structural completeness. They eliminated obvious logical errors, converted stems to a "best answer" format ("which is most appropriate?"), and reconstructed conditional links between concepts and cases to enhance formal validity.

However, AI-based refinement alone could not overcome the limitation of designing "attractive distractors," which actually dictate discrimination. In the final stage (V1.3), the human expert decisively drove psychometric completion. Specifically, sophisticated error types—such as "partial validity," "boundary cases," and "over-inference"—were strategically placed to induce incorrect responses from high-achieving students. Additionally, by readjusting cognitive levels to align with the thought processes required by each item, an assessment hierarchy ranging from literal comprehension to creative application was practically implemented.

Through this intervention, the items transitioned from tools for simple information verification to effective instruments for measuring higher-order thinking, requiring condition judgment, argument evaluation, and concept application. The uniqueness of the correct answer was also strengthened, requiring synthesis of the entire logical structure and conditions of the passage, thereby simultaneously securing discrimination and validity. Consequently, the item generation process of this study demonstrates the validity of a systematic collaborative model where AI constructs the structural and logical framework, and a human expert completes the psychometric elements. It empirically proved that in high-stakes assessments measuring higher-order thinking, ensuring distractor attractiveness and cognitive level alignment is difficult for AI to achieve independently, rendering the psychometric intervention of assessment experts indispensable. The specific refinement aspects are detailed in Table 8.

Table 8. Item generation and refinement (V1.0 → V1.3)

Category	Characteristics & Limitations of V1.0	Details of AI and Human Intervention	Characteristics & Refinements in V1.3
Distractor Design	Failure to Secure Discrimination • Distractors are overly obvious. • (Q5) Simple contradiction with the passage. • (Q6) Direct opposition to passage core; no inference needed. • (Q7) Simple antonym contrast with <Context>. • Fails to measure discrimination and critical thinking due to lack of "attractive distractors."	AI Review (V1.1~V1.2) • Self-diagnosis of distractor obviousness. • Reconstruction of "opposing viewpoints" and "complex issues" in <Context>. • Conversion of stem to "best answer" format. • Break away from simple 1:1 matching; induction of conditional connections. Human Expert Intervention (V1.3) • Points out lack of attractive distractors. • (Q5) Designs distractors applying partial validity. • (Q6) Designs distractors applying boundary cases. • (Q7) Designs distractors applying baseless over-inference.	Transition to Attractive Distractors • Shift from simple error types to distractors featuring partial validity, ov er-generalization, and over-inference. • Substantial securing of discrimination through "attractive distractors." • Enhancement of diagnostic capability for student error responses.
Cognitive Level Implementation	Superficial Implementation • Only the external form of a 4-level cognitive hierarchy is implemented. • Measurement of practical higher-order thinking processes is insufficient.	Human Expert Intervention (V1.3) • (Q8) Elevates required cognitive level from simple case application to connecting <Context> cases with the passage's "conceptual implications" (direction of institutional transition).	Alignment Based on Thought Processes • Shift from superficial classification to thought process-centric cognitive levels. • (Q6) Inference: Rule application & boundary case judgment. • (Q7) Critical: Evaluating opposing arguments & validity (over-inference). • (Q8) Creative: Expansion from case application to inferring conceptual implications.
Item Set Characteristics	Logic-Centric Items • Possesses formal structure but lacks discrimination. • Failure to control difficulty and discrimination.	Integration of Review & Design • AI's primary review of logical errors + Human's final design of psychometric quality (difficulty, discrimination, attractiveness).	Assessment-Centric Item Sets • Transition to item sets with secured psychometric quality.
Generation Method Summary	Format-Centric AI Draft • Lacks psychometric elements.	Collaborative Refinement • Integration of AI's structural correction + Human's psychometric refinement.	Development into Measurement Tools • Evolution from formal items to empirical measurement instruments. • Empirical validation of psychometric outcomes through HITL-based refinement.

Passage Quality Evaluation

This section provides a multidimensional validation of the passages generated via the HITL protocol, integrating quantitative analysis (KReaD index) and qualitative analysis (expert ratings).

Table 9. Comparison of KReaD indices between CSAT and AI-generated passages ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$)

Metric	Source	*N*	*M*	*SD*	*t*	*p*
KReaD Index	Past CSAT	15	10.879	0.523	0.367	0.747
	AI	3	10.577	1.407		
Character Count	Past CSAT	15	1707.800	445.399	2.815	0.014*
	AI	3	1383.333	13.577		
Paragraph Count	Past CSAT	15	6.000	2.478	2.310	0.036*
	AI	3	4.333	0.577		
Sentence Count	Past CSAT	15	26.067	6.239	1.144	0.295
	AI	3	23.333	3.055		
Mean Sentence Length	Past CSAT	15	15.909	1.914	1.342	0.272
	AI	3	14.370	1.794		
Sentence Structure Score	Past CSAT	15	26.165	3.817	0.654	0.565
	AI	3	24.400	4.353		
Complex Sentence Ratio	Past CSAT	15	92.638	7.650	-1.965	0.073
	AI	3	97.327	2.321		
Vocabulary Diversity Ratio	Past CSAT	15	0.454	0.095	-3.641	0.002**
	AI	3	0.550	0.017		
Difficult Vocabulary Ratio	Past CSAT	15	0.751	0.081	1.030	0.403
	AI	3	0.647	0.172		

The quantitative findings revealed no statistically significant difference in the KReaD index between the AI-generated passages (M = 10.58) and past CSAT passages (M = 10.88). Furthermore, core readability metrics—including mean sentence length, sentence structure score, complex sentence ratio, and difficult vocabulary ratio—exhibited no significant differences (Table 9). This indicates that the AI effectively reproduced the linguistic complexity and information density required for CSAT reading passages. The significantly lower character and paragraph counts in the generated passages are attributable to format differences (single vs. composite passages) rather than inherent linguistic disparities. Notably, the AI passages demonstrated a significantly higher vocabulary diversity ratio (t = -3.64, p = .002), confirming the AI's tendency to minimize redundant expressions and utilize a varied lexicon.

Multivariate similarity analysis further corroborated these results. The AI-generated passages recorded low Euclidean distances compared to the past CSAT passage clusters across distinct domains (social sciences/culture: 1.965; science/technology: 2.651; humanities/arts: 2.667). This demonstrates a high degree of multidimensional linguistic similarity, encompassing style, structure, and vocabulary distribution. Consequently, alignment with past exam passages was secured not merely on isolated metrics but at an integrated, text-level scale.

The qualitative evaluation via expert ratings indicated high structural completeness, with the AI-generated passages achieving a mean rubric total score of 32.44 out of 35 (equivalent to 4.63 on a 5-point scale). The rating data demonstrated acceptable reliability (Cronbach's α = .649, McDonald's ω = .706, mean ICC = .765), thereby substantiating the validity of the analytical findings.

An itemized analysis of the rubric revealed distinct strengths and limitations (Table 10). Criteria such as "Accuracy of Content," "Accuracy of Expression," "Unity of Theme," and "Systematicity of Structure" all scored above 4.8, demonstrating the AI's exceptional capability in formal and structural text generation. Conversely, "Assessment

Suitability" (4.44), "Clarity of Explanation" (4.35), and "Appropriateness of Difficulty" (4.22) received relatively lower ratings, with these differences being statistically significant ($p < .05$). This suggests that while AI excels at structuring text based on explicit rules, it possesses inherent limitations in fine-tuning difficulty for specific learner cognitive levels and executing educational and psychometric designs precisely aligned with assessment objectives.

Table 10. Mean scores by rubric criteria for passage quality

Rank	Evaluation Criterion	*M*	*SD*
1	2. Accuracy of Content	4.87	0.34
2	7. Accuracy of Expression	4.87	0.34
3	3. Unity of Theme	4.83	0.51
4	4. Systematicity of Structure	4.81	0.43
5	6. Assessment Suitability	4.44	0.71
6	5. Clarity of Explanation	4.35	0.76
7	1. Appropriateness of Difficulty	4.22	0.79

In summary, the AI-generated passages achieved a level of linguistic congruence and formal completeness equivalent to past CSAT passages, proving that the HITL protocol effectively reproduces the foundational text quality demanded by high-stakes assessments. However, human expert intervention remains indispensable in domains requiring educational and psychometric refinement, such as difficulty calibration, assessment suitability, and explanatory clarity. Ultimately, while generative AI significantly enhances the efficiency of passage development, an HITL-based collaborative structure is a mandatory prerequisite for securing final, high-stakes quality.

Item Quality Evaluation

This section presents an integrated validation of the generated items' quality utilizing quantitative analyses based on student response data (Classical Test Theory and Item Response Theory) and qualitative expert ratings, focusing on core findings. First, the basic CTT statistics at the item level are presented in Table 11.

Table 11. Item-level CTT statistics (difficulty, discrimination, reliability)

Passage Domain	Item	Cognitive Level	Difficulty(p)	Difficulty(p) Interpretation	Discrimination(r_{pb})	Discrimination(r_{pb}) Interpretation	Cronbach's α if Item Deleted
Humanities & Arts	Q1	Literal	0.855	Easy	0.325	Good	0.633
	Q2	Inferential	0.820	Easy	0.217	Moderate	0.647
	Q3	Critical	0.541	Moderate	0.169	Poor	0.659
	Q4	Creative	0.661	Moderate	0.394	Good	0.617
Social Sciences & Culture	Q5	Literal	0.661	Moderate	0.329	Good	0.629
	Q6	Inferential	0.444	Moderate	0.192	Poor	0.655
	Q7	Critical	0.424	Moderate	0.193	Poor	0.654
	Q8	Creative	0.596	Moderate	0.295	Good	0.636
Science & Technology	Q9	Literal	0.668	Moderate	0.407	Excellent	0.615
	Q10	Inferential	0.810	Easy	0.454	Excellent	0.612
	Q11	Critical	0.272	Difficult	0.182	Poor	0.654
	Q12	Creative	0.751	Easy	0.467	Excellent	0.606
Mean			0.625	Moderate	0.302	Good	0.656

As indicated in Table 11, item difficulty (p) ranged from .272 to .855 (M = .625), generally reflecting a moderately easy level. The mean discrimination ($r_{p\beta}$) was .302, indicating a good level overall. Test reliability (Cronbach's α) was .656, which is interpreted as above average given the small number of items and their structural heterogeneity. However, significant variance existed among items, with certain items (Q3, Q6, Q7, Q11) demonstrating poor discrimination. Q1 and Q2, moreover, slightly exceeded the target difficulty range (.20–.80), limiting their capacity to discriminate higher-ability learners.

The distribution of item characteristics is presented in the proportion-correct–discrimination scatterplot in Figure 2. The analysis revealed a positive correlation between proportion correct (p) and discrimination ($r_{p\beta}$) (r = .65, p < .05), indicating that lower-p items also tended to show lower discrimination. This pattern was driven chiefly by the critical reading items (Q3, Q7, Q11), which were both difficult and weakly discriminating. Most other items fell within an acceptable range, forming a generally valid structure overall.

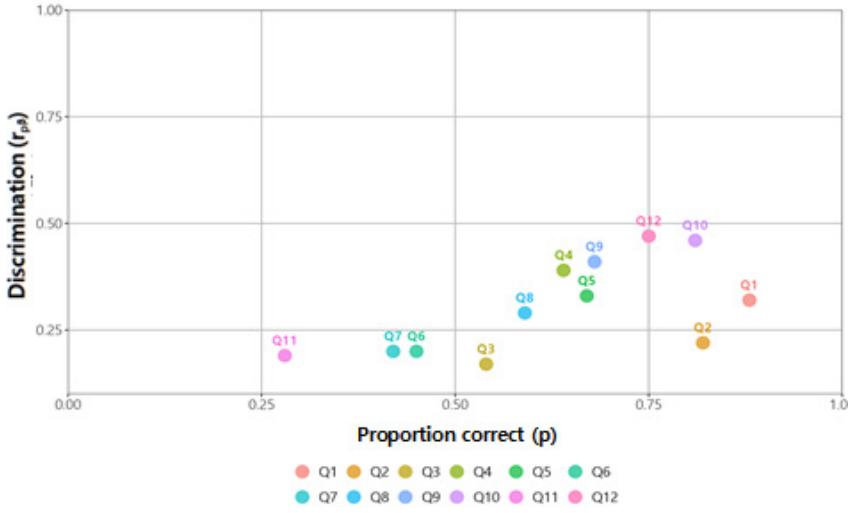

Fig. 2. Proportion-correct–discrimination scatter plot (CTT)

The analysis results by cognitive level are presented in Table 12.

Table 12. Mean difficulty and discrimination by cognitive level (CTT)

Cognitive Level	Item Numbers	Mean Difficulty(p)	Mean Discrimination($r_{p\beta}$)	Difficulty(p) Interpretation	Discrimination($r_{p\beta}$) Interpretation
Literal	Q1, Q5, Q9	.728	.354	Easy	Good
Inferential	Q2, Q6, Q10	.691	.288	Moderate	Moderate
Critical	Q3, Q7, Q11	.412	.181	Difficult	Poor
Creative	Q4, Q8, Q12	.669	.385	Moderate	Good

According to Table 12, items assessing literal and creative reading demonstrated relatively high discrimination, whereas critical reading items exhibited the lowest discrimination. This implies that critical reading items failed to effectively differentiate between respondents' ability levels. This tendency is consistently confirmed by the Item Response Theory (IRT) analysis (Table 13).

Table 13. Item parameters based on the 2PL IRT model

Passage Domain	Item	Cognitive Level	Discrimination (*a*)	Discrimination (*a*) Interpretation	Difficulty (*b*)	Difficulty (*b*) Interpretation
Humanities & Arts	Q1	Literal	1.275	Moderate	-1.777	Easy
	Q2	Inf erential	0.716	Moderate	-2.336	Very Easy
	Q3	Critical	0.492	Low	-0.354	Moderate
	Q4	Creative	1.237	Moderate	-0.696	Easy
Social Sciences & Culture	Q5	Literal	1.057	Moderate	-0.772	Easy
	Q6	Inf erential	0.494	Low	0.483	Moderate
	Q7	Critical	0.514	Low	0.634	Difficult
	Q8	Creative	0.919	Moderate	-0.498	Moderate
Science & Technology	Q9	Literal	1.450	High	-0.666	Easy
	Q 10	Inf erential	2.229	Very High	-1.110	Easy
	Q 11	Critical	0.507	Low	2.055	Very Difficult
	Q 12	Creative	2.275	Very High	-0.845	Easy
Mean			1.097	Moderate	-0.490	Moderate

The mean discrimination parameter (*a*) across all items was 1.097, an acceptable level; nevertheless, all critical reading items (Q3, Q7, Q11) remained at a "Low" level. While the difficulty parameter (*b*) generally distributed around a moderate level, some items exhibited extreme difficulty, indicating variance in measurement efficiency. Notably, Q10 ($a = 2.229$) and Q12 ($a = 2.275$) recorded "Very High" discrimination, effectively differentiating between high- and low-achieving groups.

This structure of measurement efficiency is further supported by the Test Information Function (TIF) in Figure 3. As the curve illustrates, this test provides the highest information volume in the ability level interval of $\theta = -1$ to 0, making it effective for discriminating moderate-level learners. However, information volume decreases sharply for higher ability groups ($\theta > 1$), confirming limitations in discriminating top-tier students.

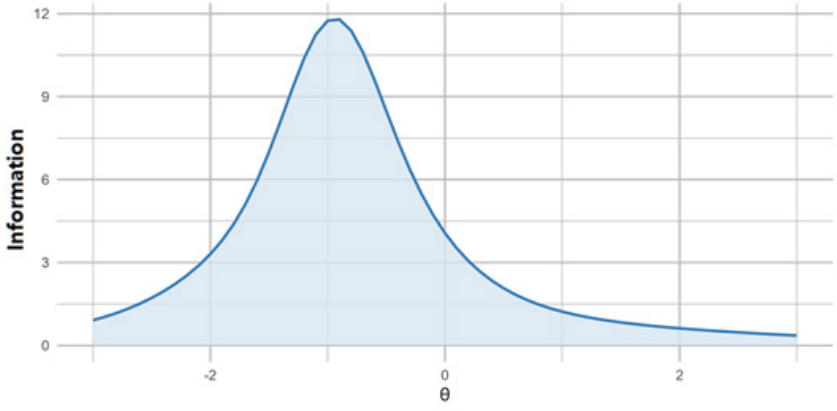

Fig. 3. Test information function (TIF)

To complement the quantitative analysis, the expert rating results are presented (Table 14). The overall mean score was 31.67 out of 35 (equivalent to 4.52 on a 5-point scale), indicating that the generated items possess a generally high level of educational completeness.

Table 14. Descriptive statistics of expert item ratings (N = 18)

Passage Domain	Item	Cognitive Level	*M*	*SD*	*SE*	95% Confidence Interval
Humanities & Arts	Q1	Literal	30.94	2.60	0.61	[29.65, 32.24]
	Q2	Inferential	32.22	2.32	0.55	[31.07, 33.37]
	Q3	Critical	32.50	2.28	0.54	[31.36, 33.63]
	Q4	Creative	32.28	2.30	0.54	[31.14, 33.42]
Social Sciences & Culture	Q5	Literal	31.28	2.59	0.61	[29.99, 32.56]
	Q6	Inferential	32.22	1.70	0.40	[31.38, 33.07]
	Q7	Critical	32.83	1.69	0.40	[31.99, 33.67]
	Q8	Creative	33.11	1.84	0.43	[32.19, 34.03]
Science & Technology	Q9	Literal	30.28	2.68	0.63	[28.95, 31.61]
	Q10	Inferential	31.11	3.46	0.82	[29.39, 32.83]
	Q11	Critical	30.56	4.36	1.03	[28.39, 32.72]
	Q12	Creative	30.72	4.34	1.02	[28.57, 32.88]
Overall Mean			31.67	2.68	0.63	

The variance across evaluation criteria is more distinctly observed in the heatmap in Figure 4. Visualizing the mean scores by color intensity reveals that "Curriculum Alignment," "Explicitness of Assessment Elements," and "Formal Completeness" received high ratings (dark colors). Conversely, "Appropriateness of Difficulty" and "Attractiveness of Distractors" received relatively low ratings (light colors), clearly exposing limitations in the domain of psychometric refinement.

In summary, the generative AI-based items generally secured a valid structure and above-average reliability in terms of difficulty and discrimination, while also receiving high ratings for curriculum alignment and formal completeness. However, the poor discrimination of critical reading items, insufficient discrimination among top-tier students, and limitations in distractor design consistently emerged. This strongly indicates the structural constraints of standalone AI generation and underscores the necessity of human expert intervention.

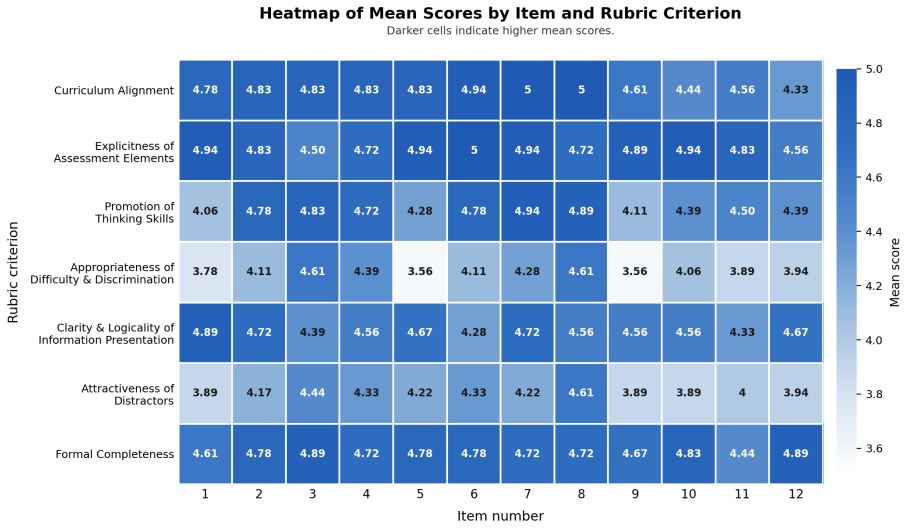

Fig. 4. Heatmap of expert rubric ratings by item (*N* = 18)

General Discussion: Effectiveness and Challenges of the HITL Protocol

This section synthesizes the quantitative and qualitative analyses of the passages and items generated via the HITL protocol to delineate the strengths and limitations of generative AI. It further establishes the teacher's role in an AIG environment and outlines directions for protocol advancement.

Initially, generative AI demonstrated distinct strengths in formal completeness and generation efficiency. In the expert ratings, both passages and items received exceptionally high evaluations across structural and normative criteria, including content accuracy, expression accuracy, curriculum alignment, and assessment element explicitness. Furthermore, readability showed no statistically significant difference from past CSAT passages. Notably, the high vocabulary diversity indicates that AI can maintain text quality while minimizing repetitive expressions. This confirms that generative AI is a robust tool capable of reliably mass-producing high-school-level reading materials.

In contrast to this formal completeness, however, a psychometric divergence emerged between empirical student response data and expert ratings. While experts evaluated the overall completeness of the items highly, empirical analysis revealed low discrimination in certain items, particularly those assessing critical reading. This implies that although experts partially recognized difficulty issues, they failed to adequately predict the extent of the actual decline in discrimination. Additionally, item quality varied by passage domain: domains with explicit conceptual structures (science and technology) exhibited high discrimination, whereas domains open to broader interpretation (social sciences/culture, humanities/arts) yielded relatively lower discrimination.

Concurrently, distractor design was identified as the decisive factor dictating discrimination. AI-generated distractors appeared superficially natural but failed to sufficiently attract high-achieving learners, thereby degrading discrimination. Critical reading items, in particular, suffered from ambiguous or complex judgment criteria, weakening the uniqueness of the correct answer and consequently failing to differentiate ability groups effectively. This gap between expert judgment and empirical function is also evident at the level of individual indices. Items that experts rated relatively high for distractor attractiveness (Q8 = 4.61, Q4 = 4.33) nonetheless yielded empirical discrimination of only .295 and .394, respectively, while the critical reading items (Q3, Q7, Q11) showed consistently low discrimination ($r_{pb} < .20$) regardless of their expert ratings. This indicates that qualitative expert judgment does not sufficiently predict an item's empirical discriminating function. Strikingly, this pattern extends to the passage level: the passage rated highest by experts for educational quality actually yielded the lowest mean item discrimination, resulting in a strong negative correlation ($\rho = -1.0$, $p < .001$) between passage rating scores and item discrimination. This demonstrates that educationally superior passages do not necessarily produce psychometrically superior items, and that holistic expert judgment possesses inherent limitations in predicting empirical measurement functions.

This limitation does not negate the necessity of human experts; it redefines their function. What failed to predict empirical performance was the expert's rating function—the holistic, post-hoc judgment of finished items—not the intervention function that actively designs discrimination and distractor attractiveness during generation. The value of the latter is evidenced directly by the V1.0-to-V1.3 refinement, in which human intervention added attractive distractors and thereby raised discrimination. In other words, the predictive failure of qualitative judgment is precisely why the expert cannot rely on it alone and must intervene through data-driven verification. Rather than serving as a mere item

author, the teacher must therefore act as an orchestrator of the generation process, intervening primarily on psychometric elements such as discrimination, difficulty, and distractor design (Burke, 2025). Furthermore, rather than relying solely on qualitative judgment, it is imperative to conduct dynamic verification based on empirical student response data. Consequently, new forms of professional expertise are required, including prompt engineering, AI utilization proficiency, and data interpretation capabilities.

Based on these findings, several directions for advancing the HITL protocol are proposed. First, differentiated prompt designs reflecting the cognitive characteristics of each reading type are required; critical reading items, in particular, demand clarified judgment criteria and sophisticated distractor design. Second, a systematic design for the timing and method of teacher intervention is necessary. An effective division of labor delegates formal reviews to the AI while concentrating human effort on psychometric refinement in the final stages. Third, establishing a documentation system that records the entire item generation process is essential to ensure transparency and reproducibility, providing a foundation for continuous improvement.

In conclusion, while generative AI can vastly improve the efficiency and formal completeness of assessment item development, the qualitative perfection of items measuring higher-order thinking is ultimately determined by the psychometric judgment of human experts. Therefore, effective item development in an AIG environment can only be realized through an HITL model that explicitly delineates and complementarily integrates the roles of AI and humans.

Conclusions and Implications

This study designed and implemented an HITL protocol specialized for the reading domain of the CSAT Korean language area, empirically validating the quality of the generated passages and items through large-scale student response data (N = 401) and expert ratings (N = 18). Analysis revealed that generative AI demonstrated high performance in formal completeness and generation efficiency for both passages and items, securing a baseline level of psychometric validity.

However, inherent limitations were identified in psychometric refinement, specifically concerning difficulty calibration, discrimination securing, and distractor design. The degradation of discrimination was particularly pronounced in critical reading items (mean a = 0.504). Furthermore, a divergence between qualitative expert judgments and empirical measurement outcomes based on student responses confirmed that educational completeness and psychometric function are not equivalent dimensions.

These findings indicate that while generative AI functions as an effective tool for formal generation and structural completeness, the finalization of the measurement function—the core of assessment—must be supplemented through human expert intervention. Accordingly, within the HITL model, the teacher must assume the role of a generation process orchestrator, focusing on psychometric judgment, data-driven verification, and distractor design.

Synthesizing the above discussions, generative AI-based AIG can substantially enhance the efficiency and scalability of assessment development. Nevertheless, a collaborative structure with human experts remains essential to secure the rigorous validity and reliability demanded by high-stakes assessments. This study is nonetheless limited by its reliance

on a single protocol implementer and a 12-item set, and by a test time allocation that did not fully replicate the actual CSAT, which may have influenced the estimates of difficulty and discrimination. Continuous research is warranted regarding the refinement of the HITL protocol and the expansion of its scope of application, along with validation under conditions that more closely approximate the actual testing environment.

Acknowledgments

This article is based on a revised version of the author's master's thesis, *AI-Based Automatic Item Generation for the CSAT Korean Reading Section,* submitted to the Graduate School of Korea National University of Education in 2026.

Conflicts of Interest

The author declared no conflicts of interest.

References

Afflerbach, P. (2017). Understanding and using reading assessment, K–12 (3rd ed.). ASCD.

Afflerbach, P., Cho, B. Y., & Kim, J. Y. (2015). Conceptualizing and assessing higher-order thinking in reading. Theory Into Practice, 54(3), 203–212. https://doi.org/10.1080/00405841.2015.1044367

Al Faraby, S., & Romadhony, A. (2024). Analysis of LLMs for educational question classification and generation. Computers and Education: Artificial Intelligence, 7, Article 100298. https://doi.org/10.1016/j.caeai.2024.100298

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

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 14437–14458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.799

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

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, Article 903077. https://doi.org/10.3389/frai.2022.903077

Baker, F. B. (2001). The basics of item response theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), Item generation for test development (pp. 199–217). Lawrence Erlbaum Associates.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Burke, C. M. (2025). AI-assisted exam variant generation: A human-in-the-loop framework for automatic item creation. Education Sciences, 15(8), Article 1029. https://doi.org/10.3390/educsci15081029

Chall, J. S. (1983). Stages of reading development. McGraw-Hill.

Cho, Y., & Lee, K. (2020). Development of the Korean language text analysis program (KReaD Index). Korea Reading Research, 56, 225–246. https://doi.org/10.17095/jrr.2020.56.8

Chung, M., Seo, S., Nam, M., Choi, S., Lee, S., & Nam, G. (2022). A qualitative analysis study on the characteristics of good Korean language assessment items: A study on the development of an item quality analysis framework for the Korean language subject (2). Journal of CheongRam Korean Language Education, 89, 43–78. https://doi.org/10.26589/jockle..89.202209.43

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. International Journal of Testing, 12(3), 273–298. https://doi.org/10.1080/15305058.2011.635830

Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. Education and Information Technologies, 29(18), 24111–24142.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hommel, B. E., Wollang, F. J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. Psychometrika, 87(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Jung, H. (2008). The actual condition of teachers' reading assessment expertise: Focusing on the analysis of written test items. Journal of Reading Research, 19, 307–346.

Kim, Y. (2015). A study on the items of the CSAT language/Korean area: Focusing on the 1994 to 2015 academic years [Unpublished doctoral dissertation]. Inha University.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Korean Reading Association. (2021). Theory and practice of reading education for reading education experts. Parkijung.

Kwon, T., Lee, J., & Kim, S. (2017). A research on the quality of reading evaluation items in College Scholastic Ability Test (CSAT) and refining way of developing items. Journal of Reading Research, 45, 131–159. https://doi.org/10.17095/JRR.2017.45.5

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. System, 123, Article 103344. https://doi.org/10.1016/j.system.2024.103344

Luecht, R. M. (2012). An introduction to assessment engineering for automatic item generation. In Automatic item generation (pp. 59–76). Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.

Nam, M., Lee, S., Choi, S., Seo, S., Nam, G., & Chung, M. (2022). A basic study on developing a quality analysis framework for Korean language assessment items. Journal of CheongRam Korean Language Education, 86, 71–95. https://doi.org/10.26589/jockle..86.202203.71

Park, G., & Choi, S. (2025a). Psychometric validity analysis of GAI-HITL-based automatic item generation (AIG) for reading comprehension items. The Journal of Curriculum Evaluation, 28(3), 319–359. https://doi.org/10.29221/jce.2025.28.3.319

Park, G., & Choi, S. (2025b). A study on automatic item generation for multiple-choice reading comprehension assessment in Korean language. Journal of Curriculum and Evaluation, 28(1), 215–246. https://doi.org/10.29221/jce.2025.28.1.215

Park, T., Park, E., Ryu, S., Han, J., Choi, S., Byun, T., Bae, H., Hong, M., Kwon, G., Lee, K., Kim, H., Oh, S., & Lee, S. (2022). Design and development of the KICE Readability Index automatic measurement program (I) (RRC 2022-9). Korea Institute for Curriculum and Evaluation.

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In Elements of adaptive testing (pp. 151–165). Springer New York.

Ryu, S. (2019). Critical examination about CSAT Korean language and its developmental directions: Toward the recovery of the nature of the CSAT evaluation. The New Korean Language Education, 121, 353–380. https://doi.org/10.15734/koed..121.201912.353

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. Educational Measurement: Issues and Practice, 43(1), 5–18. https://doi.org/10.1111/emip.12590

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024). Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), Artificial intelligence in education (pp. 165–179). Springer Nature Switzerland.

Seo, H., Lee, S., Ryu, S., Oh, E., Yun, H., Byun, K., & Pyeon, J. (2013). A study on detailing text complexity for the systematization of reading education (2). Journal of Korean Language Education Research, 47, 253–290. https://doi.org/10.20880/kler.2013..47.253

Seong, T. (2001). Understanding and application of item response theory. Kyoyookbook.

Seong, T. (2019). Modern educational evaluation (5th ed.). Hakjisa.

Seong, T., Si, G., & Choi, Y. (2024). Educational changes and the direction of educational evaluation in the era of generative AI. Journal of Educational Evaluation, 37(1), 1–28.

Shin, D. (2022). Where are the limits of AI use in English education?: The feasibility of automated text generation and automated item generation. Journal of the Korea English Education Society, 21(3), 147–163. https://doi.org/10.18649/jkees.2022.21.3.147

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

Shin, D., & Lee, J. (2024). AI-powered automated item generation for language testing. ELT Journal, 78(4), 446–452.

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://doi.org/10.64152/10125/73530

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. Interactive Learning Environments, 33(9), 5386–5405. https://doi.org/10.1080/10494820.2025.2482588

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. International Journal of Assessment Tools in Education, 12(2), 317–340. https://doi.org/10.21449/ijate.1602294

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

Wen, Z., & Chu, S. K. W. (2025). Using generative AI for reading question creation based on PIRLS 2011 framework. Cogent Education, 12(1), Article 2458653. https://doi.org/10.1080/2331186X.2025.2458653

Yang, S., Park, S., & Min, B. (2020). Development of a reading ability test tool for middle school grades 1–3 and a validation study through IRT analysis. Korean Language Education, 170, 81–122.

Yoo, Y. (2024). A study on text evaluation models using LLM-generated data [Unpublished master's thesis]. Yonsei University.

Author’s Information

Suji Lee: Korea National University of Education, Department of Korean Language Education, Doctoral Student; Cheongju Girls' High School, Teacher, First Author & Corresponding Author.https://orcid.org/0009-0008-1899-2972