STAIR-AIG: Optimizing the Automated Item Generation Process through Human-AI Collaboration for Critical Thinking Assessment

Euigyum Kim[1], Seewoo Li[2], Salah Khalil[3], and Hyo Jeong Shin[1]*
[1]Sogang University, Seoul, South Korea
[2]University of California, Los Angeles, USA
[3]MACAT International Ltd., United Kingdom

Abstract

The advent of artificial intelligence (AI) has marked a transformative era in educational measurement and evaluation, particularly in the development of assessment items. Large language models (LLMs) have emerged as promising tools for scalable automatic item generation (AIG), yet concerns remain about the validity of AI-generated items in various domains. To address this issue, we propose STAIR-AIG (*Systematic Tool for Assessment Item Review in Automatic Item Generation*), a human-in-the-loop framework that integrates expert judgment to optimize the quality of AIG items. To explore the functionality of the tool, AIG items were generated in the domain of critical thinking. Subsequently, the human expert and four OpenAI LLMs conducted a review of the AIG items. The results show that while the LLMs demonstrated high consistency in their rating of the AIG items, they exhibited a tendency towards leniency. In contrast, the human expert provided more variable and strict evaluations, identifying issues such as the irrelevance of the construct and cultural insensitivity. These findings highlight the viability of STAIR-AIG as a structured human-AI collaboration approach that facilitates rigorous item review, thus optimizing the quality of AIG items. Furthermore, STAIR-AIG enables iterative review processes and accumulates human feedback, facilitating the refinement of models and prompts. This, in turn, would establish a more reliable and comprehensive pipeline to improve AIG practices.

1 Introduction

Recent advances in natural language processing (NLP) and generative artificial intelligence (AI), particularly large language models (LLMs), have transformed educational measurement from relatively labor-intensive processes to more automated, scalable, and efficient approaches (Srinivasan, 2022; Wang et al., 2024). Prominent examples include automated scoring (Latif and Zhai, 2024; Lee et al., 2024; Luchini et al., 2025) and automated feedback generation (Hahn et al., 2021; Chan et al., 2025), which substantially improve efficiency by reducing human labor while ensuring relatively valid and consistent outcomes.

Among these innovations, automatic item generation (AIG) has emerged as a particularly pertinent application of LLM for the rapid and effective development of assessment items (Gierl and Lai, 2013; Kurdi et al., 2020). Traditional AIG approaches generated new items by replacing different numbers or words in predefined models or templates, aiming to assess the same underlying construct. With the advent of LLMs, AIG has now entered a new phase, enabling educational researchers and practitioners to generate numerous items with minimal programming expertise. However, regardless of the AIG model used, the quality, appropriateness, and validity of AI-generated items still remain questionable. Consequently, the incorporation of quality assurance processes and human participation is deemed inevitable to ensure that AIG systems are generating content as intended (von Davier and Burstein, 2024).

In particular, it is important to ensure that the assessment items are aligned with target measurement constructs, as poorly defined constructs and superficially designed items can undermine the validity and reliability of the assessment (Liu et al., 2016). Consequently, a robust human-AI collaboration (HAIC) (Fragiadakis et al., 2025) is essential not only to leverage the scalability and efficiency of the AIG process, but also to ensure overall quality and safeguard the validity of AI-generated assessment items (Hao et al., 2024). Nevertheless, prior literature reveals a lack of empirical studies validating the appropriateness of AI-generated items for assessing cognitive skills within human–AI collaborative contexts.

*Corresponding author: hshinedu@sogang.ac.kr

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 920–930

To address this gap, the present study introduces **STAIR-AIG** (*Systematic Tool for Assessment Item Review in Automatic Item Generation*), an item review tool that supports systematic and efficient human review of AI-generated assessment items. We illustrate its potential as both a practical tool and a conceptual AIG framework by applying it to the domain of critical thinking (CT), a higher-order cognitive skill widely recognized as an essential 21st-century core competency (World Economic Forum, 2015). In complex cognitive domains, such as CT, the expert review by the human is particularly important in that defining the measurement structures and developing the assessment items are quite challenging (Shin et al., 2025).

By leveraging NLP techniques, our tool provides a comprehensive linguistic feature analysis of items. This empowers human reviewers to integrate their domain knowledge in a more effective way. Furthermore, the evaluations of AIG items by human experts are stored as data, so they continuously contribute to the improvement and refinement of the internal LLMs within the AIG pipeline. In contrast to conventional methods, which generally rely exclusively on human review as a final gatekeeping measure in a linear fashion, STAIR-AIG incorporates multiple structured touch-points for expert judgment at each stage. This facilitates continuous evaluation, targeted refinement of AI-generated elements, and ongoing enhancement of LLMs for AIG through structured human feedback and prompt optimization in a dynamic manner.

In the following, we illustrate the use of the STAIR-AIG tool as a human-in-the-loop AIG process. We review the relevant literature on AIG and the traditional item review process. Then, we present a case study that demonstrates the use of the STAIR-AIG tool in the CT domain. Subsequently, we compare the evaluations performed by a human expert with those generated by the LLM to identify discrepancies and examine the implications of their collaboration for enhancing the AIG process.

2 Related Works

2.1 Automatic Item Generation

With the growing interest in AIG to build reliable computer-based assessments by stably and efficiently feeding items into the item bank, the number of publications on AIG has recently increased (Kurdi et al., 2020). Before the advent of LLMs, the techniques of AIG studies were based on syntax or templates that harness computational power to reduce human labor, such as employing grammar correction programs and developing templates to build software programs (Bejar, 1996, 2002; Singley and Bennett, 2002). In contrast, the recent rise of LLMs in the AI research field has enabled AIG researchers to generate items without extensive software engineering, while empowering item developers to effectively realize their nuanced intentions within the generation process (Attali et al., 2022; Bezirhan and von Davier, 2023).

In line with current research trends in AIG based on LLMs, this study utilizes CT items developed through a structured AIG procedure (Shin et al., 2025). This approach leverages prompt engineering techniques using LLM and is structured into three distinct modules—passage, question, and choices statements—to support systematic generation and monitoring. Within each module, detailed prompts are provided to the LLM to generate components of items intended to assess CT skills. The modules are executed sequentially to form a complete item, which is then finalized through expert review and revision. Psychometric analyses of the pilot-study data confirmed that the generated items were functioning as intended (Shin et al., 2025).

2.2 Assessment Item Review Procedure

Traditionally, the development and validation of assessment items have relied heavily on expert-driven review procedures to ensure validity, cognitive alignment, and fairness (Haladyna and Rodriguez, 2013). Guidelines from organizations such as the National Council on Measurement in Education (NCME) and the International Test Commission (ITC) emphasize the need for refinements guided by expert judgment to avoid common errors in the writing of items and to secure the validity of the construct (Haladyna and Rodriguez, 2013; Commission and of Test Publishers, 2022). However, this systematic review process, while essential, is highly time-consuming, especially in large-scale assessment contexts.

To overcome these challenges and efficiently support assessments at scale, hybrid frameworks integrating automation with human supervision are increasingly adopted. An innovative example is the *Item Factory* developed for the Duolingo English Test (DET), an item review system that incorporates human-in-the-loop processes, particularly for the development of high-stakes international DET items (von Davier et al., 2024). The *Item Factory*

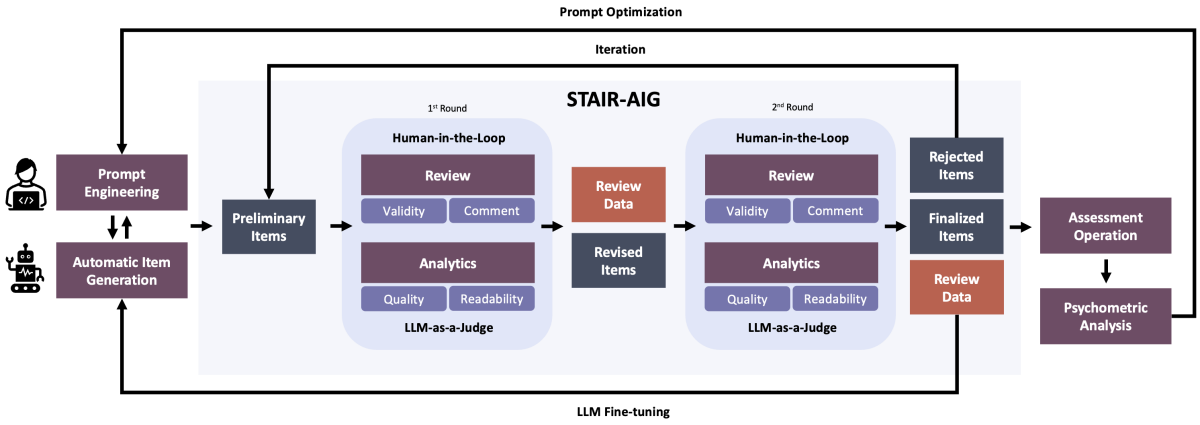

Figure 1: Pipeline of STAIR-AIG workflow

facilitates asynchronous collaboration between subject matter experts, supports reviewer calibration, and provides a structured audit trail of editorial decisions (von Davier et al., 2024). This approach not only maintains rigorous educational standards and test fairness, but also exemplifies how scalable and automated processes complemented by human oversight can enhance the quality and efficiency of assessment item review.

Item review tools, including *Item Factory*, are likely to be designed according to the types of items that are closely related to measurement constructs. To our knowledge, no open-source tool yet facilitates AIG item review for higher-order thinking skills. In the following, we present the STAIR-AIG tool and workflow as a human-in-the-loop procedure to review and optimize AIG items for CT.

3 Development of STAIR-AIG

3.1 STAIR-AIG Workflow

STAIR-AIG is developed as an iterative HAIC framework that goes beyond the static and unidirectional AIG process by continuously incorporating human reviewers' feedback to refine LLM behavior. By providing supplementary NLP features to human reviewers, human experts are expected to integrate their domain knowledge more effectively. In addition, it envisions the advancement of an AIG pipeline by automatically converting human reviews into training data for LLMs. These evaluations and human expert insights are then used to iteratively improve both AIG models through reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020) and optimize their associated prompts (Lin et al., 2024), ultimately reducing the human effort required to develop and review items that target complex cognitive constructs such as CT.

Figure 1 represents a comprehensive pipeline of the STAIR-AIG workflow. As seen in the figure, the STAIR-AIG workflow is organized as a multistage iterative loop. Preliminary items generated through prompt engineering by LLMs undergo initial evaluation and review via automated analytics, where LLMs function as auxiliary reviewers. Human reviewers then assess each item based on qualitative criteria, including content validity, appropriateness, and cognitive alignment using the STAIR-AIG tool. Importantly, reviewers provide both three-point scale ratings and open-ended feedback, and in many cases, they can directly edit the content of items. These structured data, comments, scores, and editorial changes are saved as review metadata and would be utilized to refine and enhance the performance of the AIG models.

What distinguishes STAIR-AIG is its integration of these human-generated review signals into both upstream and downstream optimization processes. On the one hand, reviewer feedback is used for prompt optimization (Lin et al., 2024), improving future item generation by refining how prompts are constructed. On the other hand, the accumulated data from reviews and edits serves as training data for RLHF (Christiano et al., 2017), fine-tuning the LLM to produce items that better align with expert judgment and the intended assessment objectives. As shown in Figure 2, this feedback loop system, inspired by the HAIC framework presented in Huang (2019), exemplifies a HAIC-based workflow designed to optimize the quality of AIG items.

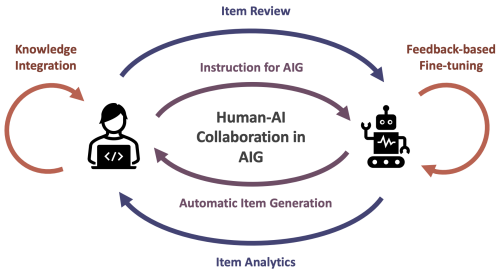

Figure 2: HAIC workflow in AIG

3.2 STAIR-AIG Modules

The STAIR-AIG system comprises two central modules designed to systematically evaluate and continuously improve the AIG process.

3.2.1 Item Analysis Module

The item analysis module operates as the preliminary review stage. Items undergo automated analysis based on quantitative linguistic metrics. The metrics include traditional NLP features, including type-token ratio, sentence length, and readability indices such as Flesch-Kincaid grade level, ensuring that the items are written clearly for the target age groups (Collins-Thompson, 2014). These metrics are selected to capture linguistic features that influence item clarity, cognitive load, and appropriateness, and to support early-stage quality screening for human review.

- **Type-Token Ratio (TTR)**: A common measure of lexical diversity, defined as

$$\text{TTR} = \frac{|V|}{|W|} \tag{1}$$

  where $|V|$ is the number of unique types and $|W|$ is the total number of tokens.

- **Average Sentence Length (ASL)**: A measure of syntactic complexity, defined as

$$\text{ASL} = \frac{N_w}{N_s} \tag{2}$$

  where N_w is the total words count and N_s is the total number of sentences.

- **Average Syllables per Word (ASW)**: A measure of word complexity, defined as

$$\text{ASW} = \frac{N_{syll}}{N_w} \tag{3}$$

  where N_{syll} is the total number of syllables and N_w is the total number of words.

- **Flesch-Kincaid Grade Level**: A readability index that estimates the school grade level required to understand a given text (Kincaid et al., 1975), calculated as

$$\text{FKGL} = 0.39 \cdot \text{ASL} + 11.8 \cdot \text{ASW} - 15.59 \tag{4}$$

We compute linguistic features by applying an XLM-RoBERTa tokenizer as a text preprocessing step (Conneau et al., 2020). Leveraging these linguistic features, the module automatically evaluates text difficulty, grade-level appropriateness, and lexical diversity metrics, which significantly reduce the workload placed upon human reviewers, thereby enhancing review efficiency and providing human reviewer with detailed item specification information to facilitate effective and timely review.

3.2.2 Item Review Module

Central to the STAIR-AIG system is the item review module, a structured interface that enables human experts to systematically evaluate AI-generated items. Items approved by the initial automated analysis are presented through this module interface. This module segments each item into specific components, such as passages, questions, and answer choices, allowing reviewers to provide detailed evaluations of each component.

Expert reviewers evaluate each component using a three-point quality scale that serves as the basis for determining whether an item would be accepted, revised, or discarded. Reviewer feedback serves a dual purpose. Qualitative comments contribute to improving the item generation prompts, while direct revision suggestions help finalize the item for operational use and also support future model refinement. Through this human-in-the-loop iterative process, STAIR-AIG continuously improves the quality and validity of the items. Once finalized, high-quality items generated by AI and modified by human experts are stored in an item bank for operational deployment. Item review module as an interface of STAIR-AIG is shown in Figure 3.

4 Empirical Research

In this empirical study, only the first round review was performed within the STAIR-AIG workflow. This initial implementation served to examine the utility of the tool and to investigate the discrepancies of review results between the human reviewer and LLM judges at the early stage of the proposed STAIR-AIG workflow.

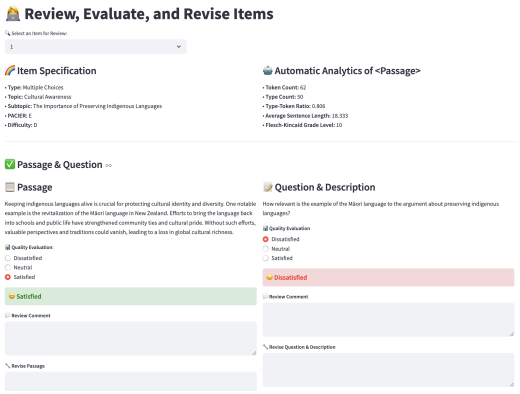

Figure 3: STAIR-AIG interface

4.1 Data

The items that were reviewed through STAIR-AIG in this study were developed by a MACAT, specializing in CT frameworks and evaluation solutions. They are based on a framework that measures and assesses CT competencies across six subdomains—Problem solving, Analysis, Creative thinking, Interpretation, Evaluation, and Reasoning (PACIER) (MACAT, 2025; Shin et al., 2025).

In this round, a total of 24 AI-generated items were reviewed, comprising multiple choice (MC) and fill-in-the-blank (FIB) types. Specifically, the assessment included 18 MC items (3 per PACIER domain) and 6 FIB items (1 per PACIER domain). Although actual CT assessment typically employs 4 choices for MC items and 3 choices for FIB items, the initial AIG items were deliberately prompted to generate 10 and 6 choices respectively, to promote a rigorous quality review without being forced to choose from all the bad choices. As for an example, an operating sample item for MACAT's CT assessment is illustrated in Figure 4.

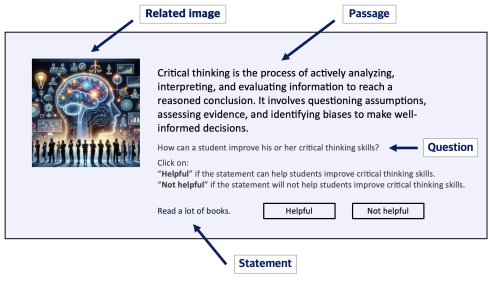

Figure 4: Sample item of CT assessment.

4.2 Item Review by Human Expert

The key review process for the 24 AIG items was conducted by a human expert who specialized in CT domain. The human expert reviewed each item systematically following the instructions and steps using the STAIR-AIG tool, indicating the quality of the items and their components on three-point rating scales.

- **Dissatisfied**: Fundamentally flawed or inappropriate item for CT assessment, and thus should be discarded. (Score: 1)
- **Neutral**: Requires revisions to improve clarity and relevance or modification of difficulty level. (Score: 2)
- **Satisfied**: Suitable for immediate use or requires minimal edits. (Score: 3)

Specifically, the expert provided ratings and comments on each of the item components, including passages, questions, choices, and overall quality of the items, referencing the analytic information provided by the item analysis module. Revision suggestions were also written directly by the expert in the open text field when necessary. Items that were rated as *neutral* or *satisfied* received detailed revision suggestions to support iterative refinement. After the review, all data including evaluations, revisions, and edits were provisionally stored as a CSV file for future model fine-tuning.

4.3 Item Quality Review by LLMs

In parallel to the human review, four OpenAI LLMs (GPT-4o, GPT-4.5-preview, o1-mini, and o3-mini) performed independent quality assessments using the LLM-as-a-judge methodology (Zheng et al., 2023). Although prior work has shown that LLM-as-a-judge is closely aligned with human preferences on a variety of tasks (Zheng et al., 2023; Gu et al., 2025), there is a lack of prior research exploring its applicability in the context of complex cognitive skills, specifically in the evaluation of the quality of the AIG items. Therefore, we explored the possibility of using LLM-as-a-judge as an additional reviewer.

Each model evaluated the AIG items based on the same criteria and the same interface used by human reviewers. The prompts were carefully aligned and mirrored with the human evaluation guidelines to ensure methodological consistency. To maintain independence between human and LLM evaluations, we adopted zero-shot learning as an in-context learning approach in which models relied solely on their pre-trained knowledge without being

provided with any task-specific examples (Brown et al., 2020). This prevents potential contamination between evaluation sources while utilizing LLM's generalized reasoning capabilities, distinct from human influence. The evaluations by LLMs were then compared with human review. Detailed prompts are provided in the Appendix A.

5 Results

5.1 Quantitative Results

5.1.1 Comparison of Human Reviews with LLM-generated Reviews

Analysis of 18 MC and 6 FIB items reveals differences in rating patterns between the human expert and LLM judges. The descriptive statistics for both item types are reported in Table 1, indicating that a human expert tends to assign lower scores overall and exhibits greater variability across all items.

In contrast, LLM judges consistently delivered higher scores across all evaluated dimensions with lower standard deviations. The o3-mini model, in particular, demonstrated extreme uniformity, assigning perfect or near-perfect scores with minimal variance. Specifically, even among LLMs, there is a subtle stratification that GPT-4.5-preview and GPT-4o exhibited slightly more variation and lower means than o3-mini. Also, in MC evaluations, the scores of the o1-mini model were closer to those of the human expert, especially in question quality.

Concretely, as illustrated in Figure 5 and Figure 6, LLMs tend to be consistently generous in their evaluations, while the human expert demonstrated a more critical and sensitive attitude marked by greater variability. A particularly notable pattern emerges in the 'Question Rating' category for FIB items, that the human expert consistently assigned the highest score to the 6 items. This uniformity is not coincidental. Since all FIB items had an identical question format, a consistent rating is justifiable and is an expected result, whereas some LLMs failed to reflect this.

5.1.2 Distribution of Ratings across Evaluators

Table 2 further illuminates the contrasting behaviors of human expert and LLM judges in evaluating the quality of AIG items. A notable pattern is the relatively frequent use of the lowest rating *Dissatisfied* (score of 1) by the human expert. Rather than indicating inconsistency, this tendency may reflect the human expert's awareness of the qualitative

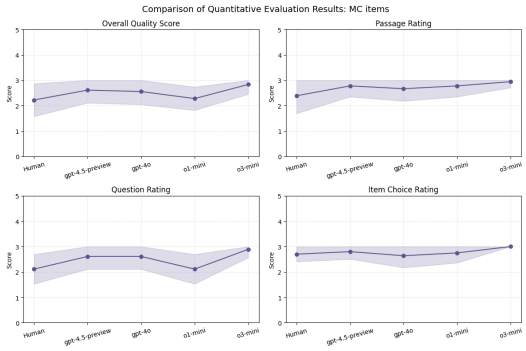

Figure 5: Rating patterns by evaluators for MC items

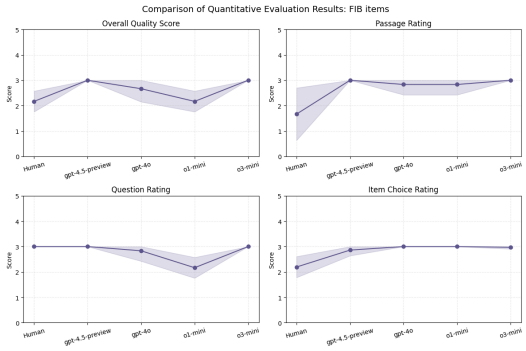

Figure 6: Rating patterns by evaluators for FIB item

aspects of the content of the item. This indicates that contextual appropriateness, coherence, and educational validity are often more readily detected through human expert, whereas automated systems may overlook such nuanced deficiencies.

In comparison, LLMs rarely gave the lowest rating of *Dissatisfied*. For example, o3-mini gave 100% *Satisfied* (score of 3) ratings in nearly every category. In the human rater effect study, this can be interpreted as a leniency or generosity (Wolfe, 2004). Even more conservative models such as o1-mini and GPT-4o showed minimal to zero use of the lowest category across MC and FIB items.

Furthermore, the human evaluator showed a more frequent use of the *Neutral* category (score of 2), which accounts for most of the responses. This middle-ground positioning can be interpreted as a nuanced case-by-case approach by the human evaluator, in contrast to the strong tendency of LLMs to assign the highest rating across most items.

5.2 Qualitative Feedback from Human Expert

To closely examine the reviews provided by the human expert, we performed a qualitative analysis of the reviewer's written comments. Table 3 lists four themes that categorize and summarize the feedback. The human expert specialized in the as-

Table 1: Descriptive statistics for MC and FIB item reviews by evaluators

Item Type	Evaluator	Overall Quality Score				Passage Rating				Question Rating				Item Choices Rating			
		Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max
MC	Human	2.22	0.65	1	3	2.39	0.70	1	3	2.11	0.58	1	3	2.70	0.29	2	3
	GPT-4.5-preview	2.61	0.50	2	3	2.78	0.43	2	3	2.61	0.50	2	3	2.80	0.30	2	3
	GPT-4o	2.56	0.51	2	3	2.67	0.49	2	3	2.61	0.50	2	3	2.60	0.54	1	3
	o1-mini	2.28	0.46	2	3	2.78	0.43	2	3	2.11	0.58	1	3	2.81	0.35	2	3
	o3-mini	2.82	0.39	2	3	2.94	0.24	2	3	2.89	0.32	2	3	3.00	0.00	3	3
FIB	Human	2.17	0.41	2	3	1.67	1.03	1	3	3.00	0.00	3	3	2.19	0.41	1	3
	GPT-4.5-preview	3.00	0.00	3	3	3.00	0.00	3	3	3.00	0.00	3	3	2.86	0.22	2	3
	GPT-4o	2.67	0.52	2	3	2.83	0.41	2	3	2.83	0.41	2	3	3.00	0.00	3	3
	o1-mini	2.17	0.41	2	3	2.83	0.41	2	3	2.17	0.41	2	3	3.00	0.00	3	3
	o3-mini	3.00	0.00	3	3	3.00	0.00	3	3	3.00	0.00	3	3	2.97	0.07	2	3

Table 2: Rating frequency and proportion for MC and FIB item reviews by evaluators

Item Type	Evaluator	Overall Quality			Passage			Question			Item Choice		
		Dissatisfied	Neutral	Satisfied	Dissatisfied	Neutral	Satisfied	Dissatisfied	Neutral	Satisfied	Dissatisfied	Neutral	Satisfied
MC	Human	2 (11%)	10 (56%)	6 (33%)	2 (11%)	7 (39%)	9 (50%)	2 (11%)	12 (67%)	4 (22%)	4 (2%)	46 (26%)	130 (72%)
	GPT-4.5	0 (0%)	7 (39%)	11 (61%)	0 (0%)	4 (22%)	14 (78%)	0 (0%)	7 (39%)	11 (61%)	3 (2%)	30 (17%)	147 (82%)
	GPT-4o	0 (0%)	8 (44%)	10 (56%)	0 (0%)	6 (33%)	12 (67%)	0 (0%)	7 (39%)	11 (61%)	28 (16%)	9 (5%)	143 (79%)
	o1-mini	0 (0%)	13 (72%)	5 (28%)	0 (0%)	4 (22%)	14 (78%)	2 (11%)	12 (67%)	4 (22%)	15 (8%)	15 (8%)	150 (83%)
	o3-mini	0 (0%)	3 (17%)	15 (83%)	0 (0%)	1 (6%)	17 (94%)	0 (0%)	2 (11%)	16 (89%)	0 (0%)	0 (0%)	180 (100%)
FIB	Human	0 (0%)	5 (83%)	1 (17%)	4 (67%)	0 (0%)	2 (33%)	0 (0%)	0 (0%)	6 (100%)	10 (28%)	9 (25%)	17 (47%)
	GPT-4.5	0 (0%)	0 (0%)	6 (100%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)	5 (14%)	31 (86%)
	GPT-4o	0 (0%)	2 (33%)	4 (67%)	0 (0%)	1 (17%)	5 (83%)	0 (0%)	1 (17%)	5 (83%)	0 (0%)	0 (0%)	36 (100%)
	o1-mini	0 (0%)	5 (83%)	1 (17%)	0 (0%)	1 (17%)	5 (83%)	0 (0%)	5 (83%)	1 (17%)	0 (0%)	0 (0%)	36 (100%)
	o3-mini	0 (0%)	0 (0%)	6 (100%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)	1 (3%)	35 (97%)

sessment of CT skills provided detailed comments, such as concerns about vague terminology, overly obvious item structure, conceptual inconsistencies, and cultural bias, which were often overlooked by LLM judges. These qualitative insights are stored as data and will play an instrumental role in shaping the future STAIR-AIG protocol, particularly in optimizing the prompts used for AIG and in systematizing the rubrics for the LLM-based review.

It is also worth noting that the human expert raised the issue of the content validity of some AIG items. Specifically, some items were on the borderline of assessing CT or reading comprehension. In such cases, the human expert not only provided a detailed explanation of their reasoning but also directly revised the wording of the items to better align with the intended purpose of the assessment. Such feedback can also be saved as data and used to fine-tune the LLMs, ultimately supporting the development of more valid and reliable AIG-powered assessment content.

6 Conclusions & Implications

6.1 Conclusions

This study introduces STAIR-AIG, a structured, human-in-the-loop framework designed to improve the quality and validity of AI-generated assessment items. Using the STAIR-AIG tool, we collected and compared item reviews from a human expert and four OpenAI LLMs. Our quantitative and qualitative analyses revealed that, while LLM's evaluations demonstrated high consistency, their feedback was generally superficial and overly lenient. Often, LLMs neglected critical issues such as ambiguous terminology, cultural insensitivity, and insufficient cognitive depth. In contrast, the human expert provided more critical and nuanced feedback, effectively identifying subtle yet significant flaws.

The two core modules of STAIR-AIG significantly support human reviewers in conducting rigorous, systematic evaluations aligned with the test-taker's background and the assessment goals, enhancing review efficiency. Notably, the discrepancies observed between human reviewers and LLM judges underscore the importance of a human-in-the-loop framework and an iterative review process. Ultimately, the data collected through these structured reviews is expected to improve the quality of AIG items and facilitate the development of more robust and refined assessment items.

6.2 Implications

As an example of a human-in-the-loop approach to AIG, this study sets the groundwork for extending STAIR-AIG into a comprehensive, full-cycle framework encompassing AIG, collaborative human-AI review, iterative refinement, pilot testing, psychometric validation, and model retrain-

Feedback Category	Review Comments
Terminology & Language Use Vague, overly technical, and structurally complex, which makes it misaligned with the assessment's purpose.	- "Do not use so many different words for the same meaning." - "(...) is a difficult formulation for not-so-strong readers." - "(...) is unnecessarily vague scientific jargon." - "The term (...) might be too technical for many students and may lead to incorrect interpretations."
Item Construction & Clue Issues Wording or structure that makes answers too obvious or misleads test-takers.	- "When mentioning acronym, use full name, and in all further mentions, use acronym." - "Change order to avoid misinterpretation." - "Answer appears verbatim in the passage." - "Too simple and easy to see the answer." - "Why use the term (...) whereas in all statements you use the term (...)? Be consistent."
Conceptual Accuracy & Fit Inaccurate or inconsistent statements, which make it unsuitable for valid assessment.	- "I have read some publications about (...), but the definition that is used here does not really fit very well." - "Biased or misleading conclusion." - "(...) and (...) depends on interpretation."
Cultural Sensitivity Culturally biased, which offers a limited perspective and potentially disadvantaging test-takers from diverse backgrounds.	- "The concept of the (...) varies by culture and perspective." - "(...) might be ideal in some contexts, while (...) may carry a clearer negative connotation." - "(...) portrayed in a one-sided positive light." - "(...) is culturally or ethically biased."

Table 3: Categorization of reviewer feedback and representative comments

ing. The human-generated reviews collected in this study would serve as a valuable resource for the first round of LLM refinement. Drawing on this empirical data, future work would focus on optimizing LLM prompting strategies and applying RLHF to improve both the quality and validity of AI-generated items. This process will help establish a more data-driven and feedback-informed basis for optimizing AIG systems.

In addition, this research contributes to the emerging field of HAIC-based test design and administration, where prior work remains limited. By demonstrating the utility of structured human reviews in guiding both AIG prompting and model fine-tuning, the study highlights a scalable pathway for the application of AI to educational measurement. Similar to how the *Item Factory* is used for DET, the proposed STAIR-AIG tool is being implemented for MACAT's CT assessment. The number of CT assessment items has rapidly doubled with the STAIR-AIG process, and the tool is being fully implemented to create an item bank of human-authored items alongside AIG for the CT assessment (Shin et al., 2024). This HAIC-driven approach showcases the increasing potential for the scholarly and sustainable use of AI in education.

6.3 Limitations

Despite its promise, this study has several limitations. First, the study was confined to an initial review by a human expert and four OpenAI LLMs, followed by a comparative analysis of their ratings. The end-to-end STAIR-AIG workflow process, particularly the integration and refinement of the AIG model through iterative review, has yet to be realized. Future work will involve more comprehensive testing of the entire STAIR-AIG pipeline.

Second, although the STAIR-AIG framework is designed to support multiple rounds of review, the current study only included one round of review by one expert reviewer. Consequently, the results may not reflect the full potential of iterative refinement, thereby limiting the framework's generalizability. Future research should explore the point at which discrepancies between LLMs and expert ratings converge. This will help us understand how LLMs behave when judging higher-order thinking skills, as well as inform the optimal stage for finalizing items for operational use and determining the maximum number of review cycles.

Third, while the item-review module was helpful to human reviewers, it could only analyze superficial metrics, such as TTR, ASL, and conventional readability indices. In the present study, grade-level suitability was judged solely based on these readability measures. Moving forward, the review module will integrate additional linguistic indicators that capture semantic dimensions in order to provide reviewers with more comprehensive support. Similarly, we did not directly measure whether the module substantially reduced the time reviewers

needed to complete their tasks. Therefore, future research would evaluate the practical effectiveness of STAIR-AIG by determining the degree to which it aids item review and the amount of time it saves compared to standard, tool-free review procedures.

Lastly, LLMs were given instructions that closely mirrored those provided to the human reviewer, yet their evaluations consistently exhibited leniency. To achieve a more harmonious integration of human and LLM ratings, future work should consider various prompt engineering techniques to calibrate LLM judgments more closely with the human evaluation standard in the CT domain. Furthermore, optimizing prompts accompanied by the psychometric results of the test data is expected to improve AIG models' ability to accurately generate and evaluate item difficulty and distractor plausibility. This would, in turn, strengthen the efficiency and validity of human-AI collaboration in test development.

Acknowledgments

This research was conducted in collaboration with the MACAT International Ltd., who provided support. We also sincerely appreciate the insightful comments and thoughtful suggestions on potential future directions for this research from the anonymous reviewers.

References

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Isaac I. Bejar. 1996. Generative response modeling: Leveraging the computer as a test delivery medium. ETS Research Report RR-96-13, Educational Testing Service, Princeton, NJ.

Isaac I. Bejar. 2002. Generative testing: From conception to implementation. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*, pages 199–218. Lawrence Erlbaum Associates, Mahwah, NJ.

Ummugul Bezirhan and Matthias von Davier. 2023. Automated reading passage generation with openai's large language model. *Computers and Education: Artificial Intelligence*, 5:100161.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sumie Chan, Noble Lo, and Alan Wong. 2025. Leveraging generative ai for enhancing university-level english writing: comparative insights on automated feedback and student engagement. *Cogent Education*, 12(1):2440182.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

International Test Commission and Association of Test Publishers. 2022. Guidelines for technology-based assessment. `https://www.intestcom.org/page/28` and `https://www.testpublishers.org/white-papers`. ISBN 979-8-88862-517-0.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2025. Evaluating human-ai collaboration: A review and methodological framework. *Preprint*, arXiv:2407.19098.

Mark J Gierl and Hollis Lai. 2013. Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32(3):36–50.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Marcelo Guerra Hahn, Silvia Margarita Baldiris Navarro, Luis De La Fuente Valentín, and Daniel Burgos. 2021. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, 9:108190–108198.

Thomas M. Haladyna and Michael C. Rodriguez. 2013. *Developing and Validating Test Items*. Routledge, London, UK.

Jiangang Hao, Alina A. von Davier, Victoria Yaneva, Susan Lottridge, Matthias von Davier, and Deborah J. Harris. 2024. Transforming assessment: The impacts and implications of large language models and generative ai. *Educational Measurement: Issues and Practice*. All authors contributed equally.

Janet Huang. 2019. Human-ai co-learning for data-driven ai. https://speakerdeck.com/janetyc/human-ai-co-learning-for-data-driven-ai. Accessed: 2025-05-03.

J. Peter Kincaid, Richard P. Fishburne, Robert L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training, U.S. Naval Air Station, Millington, TN. Archived from the original on December 10, 2020.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Xiaoqiang Lin, Zhongxiang Dai, Arun Verma, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Prompt optimization with human feedback. *Preprint*, arXiv:2405.17346.

Ou Lydia Liu, Liyang Mao, Lois Frankel, and Jun Xu. 2016. Assessing critical thinking in higher education: The heighten™ approach and preliminary validity evidence. *Assessment and Evaluation in Higher Education*, 41(5):677–694.

S. A. Luchini, N. T. Maliakkal, P. V. DiStefano, A. Laverghetta Jr., J. D. Patterson, R. E. Beaty, and R. Reiter-Palmon. 2025. Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication.

MACAT. 2025. Critical thinking assessments. https://www.macat.com/critical-thinking. Retrieved April 16, 2025.

Hyo Jeong. Shin, Seewoo. Li, Salah. Khalil, and Alina A. von Davier. 2024. Designing for adaptive testing using automatically generated items. In *Proceedings of the Annual Meeting of the International Association for Computerized Adaptive Testing (IACAT)*, Seoul, Korea.

Hyo Jeong. Shin, Seewoo. Li, Jihoon. Ryoo, Alina A. von Davier, T. Lubart, and Salah. Khalil. 2025. The nature and measure of critical thinking: The pacier framework and assessment. Manuscript submitted for publication.

Mark K. Singley and Randy E. Bennett. 2002. Item generation and beyond: Applications of schema theory to mathematics assessment. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*, pages 361–384. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Venkat Srinivasan. 2022. AI & learning: A preferred future. *Computers and Education: Artificial Intelligence*, 3:100062.

Alina A. von Davier and Jill Burstein. 2024. Ai in the assessment ecosystem: A human–centered ai perspective. In Peter Ilic, Ian Casebourne, and Rupert Wegerif, editors, *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, volume 261 of *Intelligent Systems Reference Library*. Springer, Cham.

Alina A. von Davier, Andrew Runge, Yena Park, Yigal Attali, Jacqueline Church, and Geoff LaFlair. 2024. The item factory: Intelligent automation in support of test development at scale. In *Machine Learning, Natural Language Processing, and Psychometrics*, pages 1–25. Information Age Publishing, Charlotte, NC.

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252(Part A):124167.

Edward W Wolfe. 2004. Identifying rater effects using latent trait models. *Psychology Science*, 46:35–51.

World Economic Forum. 2015. New vision for education: Unlocking the potential of technology. https://widgets.weforum.org/nve-2015/chapter1.html. Accessed April 14, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Appendix

A.1 Prompt for Item Review by LLMs

The following is an excerpt of the prompt used to instruct the LLMs in reviewing the quality of CT items. The prompt defines the evaluation criteria, output structure, and PACIER framework to assess item quality.

Listing 1: System prompt

```
You are a Critical Thinking Assessment's **Item Review
      Expert** with extensive experience in educational
      evaluation and test design, specializing in critical
      thinking.

Your role is to systematically evaluate the quality of test
      items based on established frameworks, ensuring
      fairness, reliability, and alignment with learning
      objectives.
```

Listing 2: User prompt: Review Context

```
## Review Context
- The exam items are designed for Grade 7~8 learners.
- Each item consists of a Passage, a Question, and 6 Answer
      Choices (each with an Explanation).
- Your task is to rigorously evaluate the quality of each
      component and provide structured feedback.

## PACIER Framework (Cognitive Process Dimensions)
The PACIER framework categorizes cognitive processes into six
      distinct levels:
- **Problem-Solving (P):** (...)
- **Creative Thinking (C):** (...)
- **Interpretation (I):** (...)
- **Evaluation (E):** (...)
- **Reasoning (R):** (...)
Each test item should align with at least one PACIER
      category, ensuring it assesses critical thinking skills
      effectively.
```

Listing 3: User prompt: Review Methods

```
## Evaluation Methodology
1. Assessment Criteria
   - Passage: Relevance, clarity, and cognitive demand.
   - Question: Alignment with passage, clarity, and ability
      to assess critical thinking.
   - Answer Choices: Plausibility of distractors, clarity,
      and correctness of explanations.

2. Comparative Judgment
   - Evaluate each component relative to high-quality
      reference items to ensure consistency.

3. Rating Scale
   - Dissatisfied: Fundamentally flawed or inappropriate for
      assessment and thus discarded without revision
      suggestions.
   - Neutral: Requires revisions to improve clarity,
      relevance, or difficulty. You should provide detailed
      feedback and specific revision recommendations.
   - Satisfied: Suitable for immediate use or required
      minimal edits. You could directly accept these items or
      suggest minor enhancements.

4. Actionable Feedback
   - Provide concise but specific feedback justifying each
      rating.

5. Final Output Format (Plain Key-Value Pairs, CSV-Ready)
Output only concise final results in plain key-value pairs
      (one per line) using the following CSV column structure:

Item Number, Type, Topic, Subtopic, PACIER, Difficulty,
      Overall Quality Score, Overall Comment,
Passage Comment, Passage Rating, Passage Revision,
Question Comment, Question Rating, Question Revision,
Item_1_Choice_1 Review, Item_1_Choice_1 Rating,
      Item_1_Choice_1 Revision Suggestion, ... (repeat for
      Choices 2 through 10)

## Additional Guidelines
- Ensure alignment with cognitive and linguistic proficiency
      standards.
- **Maintain consistency** across evaluations to avoid bias.
- Do not include markdown, bullet points, or additional
      explanations.
- Return only key-value pairs as output.
```