

Responsible Use of Generative AI When Creating Reading Comprehension Questions: Inference Matters

Zuowei Wang, Michael Flor, Jiyun Zu, Tenaha O'Reilly, Wanjing Anya Ma
Educational Testing Service

zwang@ets.org, mflor@ets.org, jzu@ets.org, toreilly@ets.org, anyawma08@gmail.com

Abstract

This work-in-progress paper shows how traditional item-quality indicators can miss construct-relevant differences between AI-generated and expert-created multiple-choice (MC) reading comprehension (RC) items. Items from both sources passed quality review by independent expert reviewers, and responses from 148,825 test takers showed similar item difficulty, item slope, and item fit. However, further analysis revealed differences in the distribution of required inference types. These discrepancies may escape traditional quality review but can be detected by examining whether the AI model faithfully followed prompts during intermediate item-generation steps. Evaluations of AI-generated items should therefore document prompts and generation steps to help identify and mitigate construct-relevant differences.

1 Introduction

Multiple-choice (MC) reading comprehension (RC) questions have been a staple measure of students' academic achievement for educators and researchers. The creation of high-quality MC RC questions requires expertise and effort. Recently, large language models (LLMs) have been used to generate RC items (Flor, 2025). Empirical studies have demonstrated that MC RC questions created by LLMs can have similar properties to those by human experts based on traditional indicators (Shin et al., 2025; Tomikawa & Uto, 2024). As more assessment providers use LLMs to create MC RC

questions, a critical question is whether traditional quality indicators are sufficient to ensure the quality of these items.

This work-in-progress paper compares AI-generated and expert-created multiple-choice reading comprehension (MC RC) items that had all passed the same standardized quality review process. We first examined whether the two item sources differed on traditional quality indicators, including item slope, item difficulty, item fit, and the relation between passage text complexity (Sheehan et al., 2014) and item difficulty. We then compared them on an additional dimension central to the construct of reading comprehension: the types of inferences required by the items (Cain & Oakhill, 1999; Graesser et al., 1994; Singer et al., 1992). The goal was to determine whether AI-generated and expert-created items that appeared comparable by traditional quality indicators could still differ in the inference processes they required. Because such differences may not be detected by traditional quality indicators, we argue that evaluations of AI-generated items should document the prompts and generation steps used to create them. In the session, we present a proposed approach and invite feedback on its potential to help ensure that AI-generated items function as intended.

2 Method

2.1 Sample

The study sample included 148,825 students in grades 3–12 from the United States who completed a standardized RC test in 2025. Schools used this

test to identify struggling readers and inform instruction. Periodically, the test includes a trial passage and associated items so that item parameters can be estimated before the passages become operational. Student responses to trial-passage items do not contribute to final scores, and students are not told which passage is being trialed.

2.2 Item creation

We edited 20 public-domain passages to be used as trial passages, so they had an average text complexity of grade 7 ($SD = 2$ grade levels; Sheehan et al., 2014). We randomly divided the passages into two groups of 10. For the first group, professional assessment developers wrote MC RC questions for each passage. For the second group, we used GPT-4o to generate MC RC questions using methods reported in Ma et al. (2025). Specifically, we used few-shot examples to create four types of RC MC items: gap-filling inference, text-connecting inference, pronominal bridging inference, and literal comprehension, which required little inference making.

2.3 Item quality review

Questions from both groups underwent a standardized quality review process used for high-stakes tests. The review accepted 153 MC questions. These questions were dispersed in a standardized RC test to appropriate grade levels as trial items.

2.4 Item calibration and analysis of item parameters

In this study, each trial item received responses from 3,525 to 14,740 students in the sample. Because different trial passages were used in different grade levels, and because the trial passages were administered periodically, students only responded to a subset of the items. We calibrated the items using item response theory (IRT) with a 2PL model (Embretson & Reise, 2013). Two expert-created items and one AI-generated item had very high difficulty (i.e., $b > 3$) or unusually large item slope (i.e., $a > 10$) and were removed from subsequent analysis, since they would not be included in the operational item pool after the standard item parameter review protocol of the test. This resulted in 71 AI-generated and 79 expert-created MC questions.

We compared the item parameters and item fit statistics of the two sets of remaining questions. We

also examined whether text complexity was similarly related to the difficulty of the items in both question sets.

2.5 Manual coding of items' inference types

Two human coders classified each question based on the type of inference it required (i.e., gap-filling, pronominal bridging, text-connecting, and literal). To develop a shared understanding of the coding scheme, the coders first jointly coded and discussed 18 expert-created questions from two passages. They then independently coded the remaining 142 questions from 18 passages. These independent ratings were used to calculate inter-rater agreement.

Nine expert-created MC questions did not fit the four categories because they asked students to determine the meaning of specific words in context. Both coders labeled these items as vocabulary-in-context questions. Because AI was not used to generate this item type in the present study, these questions were excluded only in the inter-rater agreement analysis and from the comparison of inference processes across the two question sets.

We examined inter-rater agreement and then compared the distribution of inference-question types across the two item sources. Further, we examined the few-shot prompts used to generate the questions and analyzed to what extent the generated questions conformed with the corresponding prompt.

3 Results

3.1 AI-generated and expert-created items show similar item parameters.

The expert-created and AI-generated MC questions had similar item parameters and item fit statistics, including slope, difficulty, and RMSEA – we found no significant difference between expert-created and AI-generated items on slope, difficulty, or RMSEA. In both sets of items, passage text complexity (Sheehan et al., 2014) predicted item difficulty in a similar way; there was no interaction between item source and text complexity. Table 1 shows the detailed results.

	Expert	GPT-4o
N (items)	79	71
N (passages)	10	10
Mean (Text Complexity)	707	706
SD (Text Complexity)	205	172
Range (Slope)	[.36, 2.59]	[.36, 2.10]
Mean (Slope)	1.31	1.26
SD (Slope)	0.44	0.44
Range (Difficulty)	[-.82, 2.33]	[-.86, 1.88]
Mean (Difficulty)	0.01	0.03
SD (Difficulty)	0.59	0.59
Range (RMSEA)	[0, 0.01]	[0, 0.02]
Mean (RMSEA)	0.006	0.006
SD (RMSEA)	0.003	0.003

Table 1: Distribution of Text Complexity and Item Parameters by Source

3.2 Inter-rater reliability when coding inference type of items

When coding the type of inference of items, the two human coders' inter-rater reliability as measured by Cohen's Kappa is .76 for the 52 independently coded expert-created items and .76 for the 71 independently coded AI-generated items, suggesting substantial agreement.

	AI-Generated	Expert-Created
Gap-Filling	9 (18%)	20 (42%)
Pronominal Bridging	30 (59%)	19 (40%)
Text-Connecting	12 (24%)	9 (19%)
Total	51 (100%)	48 (100%)

Table 3: Number of questions by inference type and source

3.3 AI-generated and expert-created items show different inference type distributions.

Forty-eight (61%) of the 79 expert-created items and 51 (72%) of the 71 AI-generated items required inference making. The manual coding revealed differences in the distribution of inference-question types across item sources. Specifically, AI tended

	Gap-Filling	Pronominal Bridging	Text-Connecting
Gap-filling	9	0	0
Pronominal Bridging	7	15	8
Text-Connecting	3	1	8
Literal	8	5	7

Table 2: Inference questions: target (columns) vs. actual (rows)

to create more pronominal bridging and text-connecting questions but fewer gap-filling questions, $\chi^2(2) = 6.99$, $p < .05$. Table 2 shows the different distributions of inference types across item sources.

3.4 The AI model often failed to generate the target inference question specified by the prompt.

Further analysis of LLM's item generation process revealed that it often failed to produce specific types of questions as instructed in the prompt. For example, when the prompt asked the LLM to create gap-filling questions, only 9 of the 27 questions created were actually gap-filling questions, with 7 coded as pronominal bridging, 3 coded as text-connecting, and 8 as literal questions that did not require inference making (Table 3). Overall, only 45% of the questions generated by the LLM were consistent with the target inference type specified by the prompt. This failure to create specific types of inference questions appears to explain the different composition of item types in the expert-created vs. AI-generated questions.

4 Discussion

The standard item review process adopted by the RC test did not examine whether questions required the intended types of inference. If we had relied only on the usual quality review process, the difference in the distribution of inference types between expert-created and AI-generated MC RC questions would have remained hidden. Because inference making is central to reading comprehension, even subtle shifts in the kinds of inferences elicited by AI-generated items could affect what the test measures. The present findings

therefore suggest that AI-generated RC items may contain important construct-relevant differences that are not captured by traditional quality indicators alone.

This study shows that evaluating AI-generated RC questions requires looking beyond final item statistics and reviewer judgments. Here, the AI-generated items met predetermined standards and showed comparable psychometric properties, yet they did not consistently reflect the inference-question types specified in the prompt. This mismatch became visible only when we examined the intermediate generation process. We therefore argue that responsible evaluation of AI-generated RC questions should include transparent documentation of prompts, generation steps, and verifications of whether generated items conform to the intended design.

This kind of process-level evaluation also creates opportunities for mitigation. Once prompt-following failures are identified, developers can adjust prompts, add targeted review criteria, or intentionally generate additional underrepresented item types. For example, if LLMs underproduce valid gap-filling inference questions, developers could generate a larger pool of such questions and review them specifically for inference quality. In this way, transparency about the generation process is not only a safeguard against hidden weaknesses but also a practical path toward improving AI-facilitated item development.

References

- Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5-6), 489-503.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Flor, M. (2025). *Automatic question generation*. Springer.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review, 101*(3), 371. <https://doi.org/10.1037/0033-295X.101.3.371>
- Ma, W. A., Flor, M., & Wang, Z. (2025). Automatic generation of inference making questions for reading comprehension assessments. *arXiv preprint arXiv:2506.08260*.
- Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal, 115*(2), 184-209. <https://doi.org/10.1086/678294>
- Shin, D., Kwon, S. K., & Lee, Y. (2025). Examining the efficacy of generative artificial intelligence in item generation: Comparative analysis of human-developed and ai-generated reading tests. *Education and Information Technologies, 30*(16), 23981-24007. [10.1007/s10639-025-13683-6](https://doi.org/10.1007/s10639-025-13683-6)
- Singer, M., Andruslak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. *Memory & Cognition, 20*(5), 539-548. [10.3758/BF03199586](https://doi.org/10.3758/BF03199586)
- Tomikawa, Y., & Uto, M. (2024). Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt, *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky Cham*.