Comparing AI tools and human raters in predicting reading item difficulty

Hongli Li[1], Roula Aldib[2], Chad Marchong[3], Kevin Fan[4]

[1, 3] Department of Educational Policy Studies
[2] Department of Psychology
Georgia State University
Atlanta, GA.

Abstract

This study examines how well generative AI can predict the difficulty level of reading comprehension items. Four AI tools (i.e., ChatGPT-5, Claude Sonnet 4, Gemini 2.5 Pro, and DeepSeek R1) were evaluated alongside two human raters on 20 items. Claude and Gemini showed the highest agreement with empirical values, in some cases matching or surpassing human raters, while ChatGPT-5 and DeepSeek performed less well. All AI tools and human raters tended to overestimate item ease, though Claude provided the most accurate estimates. These findings suggest that AI tools can complement expert judgment in test development, but empirical validation is necessary for ensuring accurate difficulty estimation.

1 Introduction

In traditional test development, the early stages typically involve field testing to gather pilot data to conduct item analysis. Based on the preliminary results, test items may be revised or discarded as necessary to improve the assessment quality. This process of data collection and analysis is often costly and time-consuming. While generative AI is increasingly recognized for assisting with test development (Bezirhan and von Davier, 2023; Dueñas et al., 2023), its capacity to evaluate item psychometric properties, such as item difficulty, during test development remains unclear.

According to classical test theory (CTT), item difficulty refers to the proportion of test takers who answer an item correctly, known as the p value. Higher p values indicate easier items, while lower p values correspond to more difficult items. A recent study (Li and Marchong, 2024) used ChatGPT to estimate item difficulty for a reading comprehension test and reported moderate correlations ($r > .40$) between ChatGPT's predictions and empirically derived difficulty values. However, that study did not include a direct comparison between AI-generated predictions and human estimates. It remains unclear how AI estimates compare to those made by human experts. Also, it is unknown which AI tools are better suited than others for this task.

To address these gaps, we expand on our previous work by incorporating multiple AI tools and human raters. This study aims to evaluate the effectiveness of AI tools in predicting the difficulty of reading comprehension items in comparison to human raters, with empirical item difficulty (derived from CTT) as the benchmark.

2 Literature Review

Several factors have been systematically identified as influencing reading item difficulty. For instance, Davey (1988) examined a wide range of factors that may contribute to item difficulty, including passage variables (e.g., length, coherence, and syntactic complexity), question types (e.g., response location, inference type), and question format (e.g., stem characteristics and distractor plausibility). Their regression analysis found that stem length and location of response information accounted for a significant amount of variance in item difficulty. Lumley et al. (2012) specifically identified ten factors to predict PISA reading item difficulty, including the number of features and conditions to be comprehended, proximity of pieces of required information, competing information, prominence of necessary textual information, relationship between task and required information, semantic match between task and text, concreteness of information, familiarity of information needed to answer the question, register of the text, and extent to which information from outside of the text is required to answer the question. Overall, studies have shown that both

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 84–89

passage features (e.g., vocabulary sophistication, readability, discourse cohesion) and item features (e.g., stem length, response format, and distractor quality) are associated with reading item difficulty (Choi and Moon, 2020; Davey, 1988; Lumley et al., 2012; Rafatbakhsh and Ahmadi, 2023).

In addition, expert judgment has traditionally been used to estimate item difficulty, often as part of standard-setting or early test development. According to a systematic review conducted by Alkhuzaey et al. (2024), 34% of the included studies compared their systems' predictions with experts' judgement, and on average three experts were recruited per study to judge item difficulty. However, research indicates that expert ratings are subject to bias and inconsistency. For example, Sayın and Bulut (2024) found that although expert predictions of reading item difficulty improved after feedback, their initial ratings often diverged from empirical results. The procedures for expert judgment also vary considerably across studies. In some cases, training was not provided to experts (e.g., Choi and Moon, 2020), or criteria for evaluating difficulty were not clearly specified (e.g., Davey, 1988; Desai and Moldovan, 2019). Such variability raises concerns about the reliability of expert judgments (Alkhuzaey et al., 2024). In sum, while expert review remains common in test development, judgments of item difficulty are often inconsistent and imprecise.

Recently, researchers have begun to explore whether artificial intelligence can provide more consistent predictions of item difficulty than traditional methods. For example, Li, Jiao, and colleagues (2025) modeled item difficulty in large-scale assessments using both small and large language models with different data augmentation strategies. They reported that GPT-4 did not perform as strongly as expected, likely due to limited training data, and suggested that additional data or more advanced reasoning techniques may be required. Their work was based on data from National Board of Medical Examiners (NBME) and the items were about medical practice. This raises the question of whether similar findings extend to reading assessments. It is also unclear whether AI offers advantages over human judgment or simply mirrors its limitations.

Therefore, in this study, we focus on reading comprehension items to investigate whether AI tools (especially LLMs) can provide accurate and reliable estimates of item difficulty. Specifically, we compare predictions from multiple AI tools with human expert ratings and with empirical values derived from examinee responses.

3 Methods

3.1 Instruments and Participants

The reading comprehension test used in this study was a reading section of an English proficiency test. This test evaluates advanced level English language competence of adult non-native speakers of English who plan to use English for academic purposes in a university setting. This test assesses examinees' understanding of college-level reading texts and includes four passages, each followed by five multiple-choice items, for a total of 20 items. Each item has four options, including one answer key and three distractors. All four passages were adapted from newspaper articles. Empirical response data are available from a sample of 2,019 examinees.

The AI tools tested included ChatGPT-5, Claude Sonnet 4, Gemini 2.5 Pro, and DeepSeek R1. ChatGPT-5 is a multimodal model with improved reasoning compared to earlier versions. Claude Sonnet 4 is a medium-sized model with extended context capacity, developed for reasoning and code-related tasks. Gemini 2.5 Pro is also multimodal, with enhanced long-context processing. DeepSeek R1 is an open-source model trained with reinforcement learning, designed to balance reasoning performance with computational efficiency.

Two human raters were invited to provide ratings as well. One was a non-native English speaker with extensive experience teaching English to ESL learners. The other was a non-native English speaker who held a graduate degree and had some experience in literacy research and ESL instruction.

3.2 Data Collection Procedures

Four AI tools were asked to estimate the difficulty of the 20 items on August 30th, 2025. As shown in Appendix A, the authors provided a rating form based on a thorough review of the literature. Below are the sample prompts used to interact with the AI tools:

Researcher: *I have attached a document "rating direction" where you can see the direction of the*

task. I've also provided the document 'Text" which includes the reading comprehension test. Do you understand the task?

AI tool: *Yes, I understand the task….*

Researcher: *Now I am going to give you a slightly different task. Instead of giving the 5-category rating, can you provide a more nuanced estimation of item difficulty as p value in the classical test theory. Do you understand the task?*

AI tool: *Yes, I understand the task…*

As a result, each of the four AI tools generated both categorical ratings on a 1–5 scale and continuous ratings on a 0–1 scale for each of the 20 reading comprehension items.

In parallel, two human raters independently evaluated item difficulty on a 1–5 scale using the same provided materials; however, they were not asked to provide continuous ratings. To establish a benchmark, empirical difficulty values (p values) were calculated for each item based on the responses of 2,019 examinees.

4 Results

The predictions from both AI tools and human raters were compared to the empirical values with different approaches. First, we used Spearman correlations to examine the association between the estimated categorical difficulty ratings and the empirical p values. The original categorical ratings were coded as 1 = easiest and 5 = hardest. To align the direction of the scales (since higher p values indicate easier items), categorical ratings were reverse coded before correlation analyses. As shown in Figure 1, among the AI tools, Claude demonstrated the strongest alignment with empirical values (ρ = .66), followed by Gemini (ρ = .52) and ChatGPT (ρ = .41). DeepSeek showed the weakest association (ρ = .19). For the human raters, Rater 2 exhibited moderate alignment with empirical values (ρ = .52), while Rater 1 showed lower consistency (ρ = .29). These results suggest that certain AI tools, particularly Claude and Gemini, can approximate empirical item difficulty as well as or better than human raters.

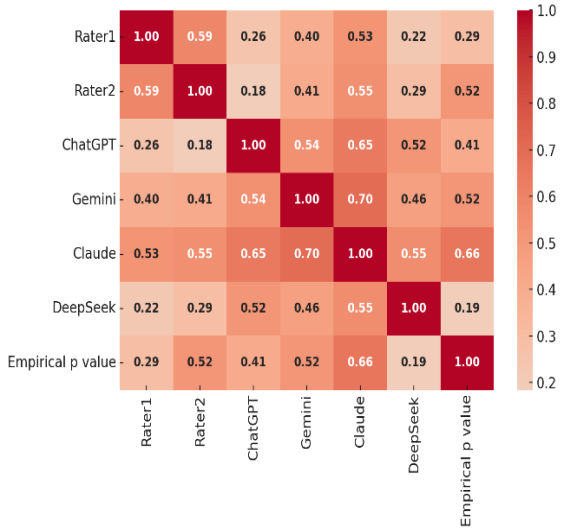

Figure 1: Spearman correlations among categorical ratings by human raters, AI tools, and empirical p value.

We also calculated quadratic weighted kappa (QWK) to evaluate agreement among the two human raters and four AI tools. QWK was selected because it accounts for the ordinal nature of the 1–5 scale and differentially weights disagreements based on their magnitude. As shown in Figure 2, agreement between the two human raters was moderate (κ = .54). Among the AI tools, Claude showed the strongest consistency with both human raters (κ = .44–.47) and other AI tools (κ = .68 with Gemini and κ = .65 with ChatGPT-5). Gemini also demonstrated strong agreement with ChatGPT-5 (κ = .61) and Claude (κ = .68). In contrast, DeepSeek exhibited only moderate agreement with both humans (κ = .24–.32) and the other AI systems (κ = .41–.57). Overall, Claude and Gemini not only aligned most closely with empirical difficulty values but also achieved the highest inter-rater consistency, while DeepSeek showed weaker agreement with others.

Using the AI tools' continuous 0–1 difficulty ratings, we calculated Pearson correlations with the empirical p values. Results mirrored the categorical analysis: Claude showed the strongest association (r = .60), followed by Gemini (r = .57), ChatGPT-5 (r = .43), and DeepSeek (r = .20).

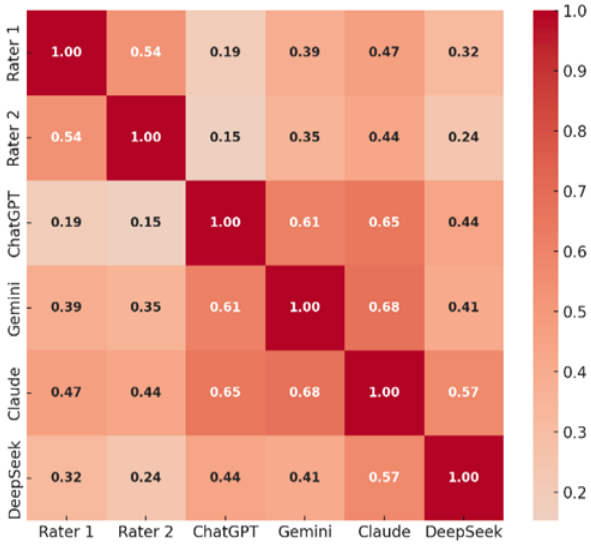

Figure 2: Quadratic weighted kappa agreement between categorical ratings by human raters and AI tools.

Furthermore, as shown in Figure 3, all four AI models systematically overestimated item ease (positive bias). We, therefore, calculated Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between empirical p values and continuous ratings by AI tools. MAE was calculated by averaging the absolute value of the errors, which indicates the average size of the deviations regardless of direction. RMSE was calculated by taking the square root of the averaged squared errors, which is more sensitive to occasional large discrepancies.

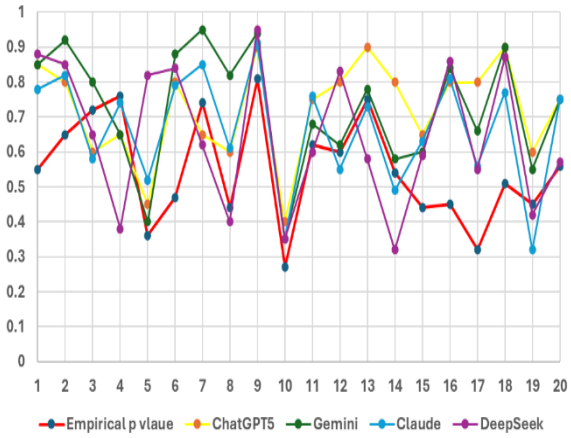

Figure 3: Comparison of continuous ratings by AI tools and empirical p values.

As shown in Table 1, in terms of error magnitude, Claude yielded the lowest mean absolute error (MAE = 0.157) and root mean square error (RMSE = 0.181), indicating the closest alignment with empirical p values. This was followed by Gemini (MAE = 0.186, RMSE = 0.230) and ChatGPT-5 (MAE = 0.204, RMSE = 0.231). DeepSeek R1 showed slightly lower MAE than ChatGPT-5 (0.201 vs 0.204) but the highest RMSE (.244), indicating more large-error outliers. These results suggest that although all AI tools tended to rate items as easier than observed in empirical data, Claude provided the most accurate continuous predictions.

AI Tools	MAE	RMSE
ChatGPT-5	0.204	0.231
Gemini 2.5 Pro	0.186	0.230
Claude Sonnet 4	0.157	0.181
DeepSeek R1	0.201	0.244

Table 1: Error magnitude of AI rating against empirical p values.

5 Discussion

In this study, we found that certain AI tools, particularly Claude and Gemini, can approximate empirical item difficulty as well as, or in some cases better than, human raters. This suggests that AI tools could serve as a valuable supplement to expert ratings in this task. In Li and Marchong (2024), only ChatGPT and the OpenAI API were used to estimate the difficulty of the same 20 reading comprehension items. Their reported Pearson correlations with empirical p values were .48 for ChatGPT-4o and .29 for ChatGPT-4o mini. In the present study, the Pearson correlation between ChatGPT-5 and empirical p values was .43, indicating that ChatGPT has not demonstrated noticeable improvement in predicting reading item difficulty over the past year. By contrast, Claude achieved the highest correlation with empirical p values (r = .60), followed by Gemini (r = .57). These findings suggest that Claude and Gemini currently offer more promising performance than ChatGPT for estimating reading item difficulty.

Our results echo the findings in Li, Jiao and colleagues (2025), who reported that GPT-4 showed limited performance in estimating difficulty of medical practice items, likely due to limited training data. Notably, the RMSE values for GPT-4 in their study (> .35) were higher than those observed in ours, where the RMSE for ChatGPT-5 was .231. This may reflect domain differences. Overall, their findings in medicine and ours in reading comprehension suggest that while LLMs

show promise for predicting item difficulty, their effectiveness may depend heavily on model design, training, and the assessment.

Furthermore, in this study, both human raters and AI tools systematically rated items as easier than indicated by the empirical p values. Human experts are prone to underestimating how challenging items are for less proficient examinees, because they are much more proficient than examinees (Nathan and Petrosino, 2003). It seems that AI tools had the same tendency, maybe they are likely to perceive items as easier given their own massive intelligence. In addition, both humans and AI tools may have underweighted the role of distractors in multiple-choice items, which often contribute substantially to empirical difficulty. These findings show the necessity of complementing expert or AI-based predictions with empirical validation.

6 Conclusion

By comparing AI-based predictions to both human expert judgment and empirical values, we aim to understand whether AI models can reliably contribute to the early-stage evaluation of test items. Our findings show that Claude and Gemini achieved the highest agreement with empirical values, outperforming ChatGPT-5 and DeepSeek. Both Claude and Gemini also demonstrated stronger alignment with human raters and were able to predict item difficulty as well as, or in some cases better than, human raters.

However, the four AI tools as well as the two human raters systematically overestimated item ease, though error analyses (MAE, RMSE) suggested Claude provided the most accurate estimates. These findings indicate the potential of AI tools to supplement human judgment in test development; at the same time, they also show the need to include empirical evidence to cross-validate AI-based difficulty estimation.

7 Limitations

While this study provides important insights on the potential of generative AI to support reading assessment development, it also has several limitations. First, the analysis was based on only 20 reading comprehension items drawn from four expository passages. Thus, the findings may not generalize to other item types, genres, or reading assessments. Second, only two human raters were included. A larger pool of experts, possibly with varied backgrounds (e.g., item writers, teachers, researchers), could provide a more reliable benchmark of human judgment. Third, while we used the default outputs of four AI tools, future research could examine how different prompting strategies or fine-tuning approaches influence prediction accuracy.

References

Alkhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914. https://doi.org/10.1007/s40593-023-00362-1

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

Choi, I. C., and Moon, Y. 2020. Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, 17(1):18–42. https://doi.org/10.1080/15434303.2019.1674315.

Davey, B. 1988. Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *The Journal of Experimental Education*, 56(2):67–76. https://www.jstor.org/stable/20151717.

Desai, T., and Moldovan, D. I. 2019. Towards predicting difficulty of reading comprehension questions. In *Proceedings of the Thirty-Second Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pages 8–13. https://cdn.aaai.org/ocs/18267/18267-78886-1-PB.pdf

Dueñas, G., Jimenez, S., and Ferro, G. M. 2023. You've got a friend in... a language model? a comparison of explanations of multiple-choice items of reading comprehension between ChatGPT and humans. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 372–381, Toronto, Canada. Association for Computational

Linguistics. https://aclanthology.org/2023.bea-1.30/

Li, H., and Marchong, C. 2024. Evaluating item difficulty of a reading comprehension test using ChatGPT. Paper presented at the GSU Adult Literacy Research Center (ALRC) Mini-Conference, Atlanta, GA.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., and Lissitz, R. W. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644251344973

Lumley, T., Routitsky, A., Mendelovits, J., and Ramalingam, D. 2012. *A framework for predicting item difficulty in reading tests*. Australian Council for Educational Research (ACER). https://research.acer.edu.au/pisa/5/.

Nathan, M. J., and Petrosino, A. 2003. Expert blind spot among preservice teachers. *American Educational Research Journal*, 40(4):905–928. https://doi.org/10.3102/00028312040004905

Rafatbakhsh, E., and Ahmadi, A. 2023. Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):41. https://doi.org/10.1186/s40862-023-00214-4

Sayın, A., and Bulut, O. 2024. The difference between estimated and perceived item difficulty: An empirical study. *International Journal of Assessment Tools in Education*, 11(2):368–387. https://doi.org/10.21449/ijate.1376160

Appendix A: Rating Form

Purpose of the Task
We are conducting a study to estimate the difficulty of 20 reading comprehension items. The test is designed for advanced adult nonnative English speakers preparing for academic study, and measures their ability to understand college-level texts. It consists of four passages adapted from newspaper articles, each followed by five multiple-choice items, for a total of 20 items.

Item difficulty here refers to how likely it is that an average member of the target group (advanced adult non-native English speakers) will answer the item correctly.

Factors You May Consider

1. Linguistic Features
These are characteristics of the reading text itself:
- Vocabulary – Rare words, technical terms, or high lexical density (lots of content words) make comprehension harder.
- Sentence Structure – Longer sentences, more clauses, and complex syntax increase difficulty.

2. Item Features
These are aspects of the test item itself:
- Question Type –
  - *Literal* (answer explicitly in the passage) = easier.
  - *Inference* (paraphrase, bridging, gist, or prior knowledge) = harder.
- Question Phrasing – Negatively worded or confusing stems add difficulty.
- Item Length – Long stems or long answer options increase processing load.
- Answer Options –Longer or more complex distractors make the question harder.

3. Cognitive Demands
These relate to the mental processes required:
- Locating Explicit Information – Easier (requires simple scanning).
- Integrating Across Sentences – Moderate difficulty (requires synthesis of information).
- Higher-Level Inference or Reasoning – Hardest (requires abstraction, generalization, or drawing on prior knowledge).

Please use the provided features holistically, and rely on your expert judgment, experience, and instinct. The goal is to provide your expert impression of relative difficulty.

Rating Scale (5-Point)
Please assign one rating (1–5) for each item:
1. Very Easy – Almost all test-takers are expected to answer correctly.
2. Easy – Most test-takers are expected to answer correctly.
3. Moderate – About half of test-takers are expected to answer correctly.
4. Difficult – Fewer than half of test-takers are expected to answer correctly.
5. Very Difficult – Only a small proportion of test-takers are expected to answer correctly.

Note. Please bear in mind that the target test-takers are advanced non-native English speakers who are seeking to study at English speaking institutions.