

# Using Generative AI to Simulate Item Responses by Skills Insight Score Band

Jianshen Chen<sup>1</sup>, Chansoon Lee<sup>2</sup>, Kylie Gorney<sup>3</sup>, Judit Antal<sup>1</sup>

<sup>1</sup>College Board, <sup>2</sup>American Board of Internal Medicine, <sup>3</sup>Michigan State University

## Abstract

This study evaluates GPT-5.4 Thinking simulations of reading-item responses using Skills Insight score-band descriptions and item content. Simulated and empirical item statistics and IRT parameters were compared. Difficulty recovery was strongest, with moderately strong b-parameter correlations, while a and c recovery was weaker, supporting preliminary difficulty evaluation before field testing.

## 1 Introduction

Field testing has traditionally been used to evaluate the psychometric properties of newly developed items, including difficulty, discrimination, and potential bias. However, field testing is time-consuming, costly, and resource-intensive, motivating interest in alternative approaches.

Text-based methods have been used to predict item difficulty (for a recent review, see Peters et al., 2025) and other item properties. Although promising, these methods produce predicted statistics, such as classical item statistics and item response theory (IRT) parameters, rather than response-level data. More recently, researchers have used large language models (LLMs) to simulate item responses (e.g., Benedetto et al., 2024; Demirkaya et al., 2025; Maeda, 2025; Tang et al., 2026). Because simulated data can be analyzed similarly to field-test data, they may provide richer information before operational use.

## 2 Purpose of the Study

When simulating item responses, the ability distribution should reflect the target population. This study uses Skills Insight, developed by

content specialists, which describes the knowledge, skills, and abilities (KSAs) expected at different score bands. Unlike prior studies that use generic student levels (e.g., Benedetto et al., 2024), IRT ability levels (Demirkaya et al., 2025; Maeda, 2025), or cognitive diagnosis model mastery profiles (Tang et al., 2026), Skills Insight score bands describe what students in specific score ranges are likely to know and do. This allows evaluation of ordered performance across score bands, alignment between item correctness, rationales, and expected skills, and preservation of item- and score-level patterns.

This study investigates whether LLMs can emulate examinees across score bands by generating item responses and brief rationales consistent with Skills Insight KSAs and comparable to operational data. The goal is to evaluate whether Skills Insight-anchored and LLM-generated response data can reproduce empirical patterns in classical item statistics and IRT parameter estimates. If agreement is acceptable, content specialists could use LLMs to obtain early indications of item difficulty and likely responses across score bands.

## 3 Simulation Study

### 3.1 Data

Three sources of information are provided to GPT:

1. Skills Insight KSA descriptions for seven score bands on a reading assessment.
2. An empirical ability distribution that specifies the proportion of examinees within each of the seven score bands.
3. Item content, including stems, response options, answer keys, and expert-assigned content categories.

The study uses 50 reading items from a fixed form.

### 3.2 Methods

GPT-5.4 Thinking (OpenAI, 2026) is used to predict whether each simulated student in a specific Skills Insight band is likely to answer each item correctly and generate item scores (i.e., 0 for incorrect or 1 for correct). For each run, the model allocates students across the seven score bands and predicts item responses based on the alignment between cumulative band-level KSAs and the cognitive and linguistic demands implied by item content and expert-labeled categories. The sample prompt and output are shown in Appendices A and B, respectively.

Simulations are conducted under three sample sizes per item:  $N = 500, 1,000$ , and  $2,000$ . Each sample size is run with 10 independent replications using the same prompt and input files. Two custom GPTs are created: one without training and one with training. Note that in this study, “training” refers to task-specific refinement using empirical item information. The underlying GPT model parameters are not updated. The 50 items from a fixed form were split into a training set of 30 items and a validation set of 20 items. In the without training scenario, GPT uses the provided files and prompt (see Appendix A) to generate responses and rationales for the 20 items in the validation set directly. In the training scenario, for the 30 training items, empirical IRT parameters are provided, and GPT compares the content-based predicted probability of a correct response with the IRT parameter-based probability. When the absolute discrepancy exceeds a user-set criterion (i.e., 0.20 in this study), GPT refines its content model and prediction. The refined content-based approach is then applied to the 20 validation items, for which the empirical IRT parameters are not provided to GPT. The same prompt shown in Appendix A is used for generating predicted item responses of the 20 validation items.

Classical item statistics, including proportion correct ( $p$ ) and point-biserial correlation ( $r_{pbis}$ ), are calculated from GPT-predicted responses, and IRT parameters are estimated based on GPT-predicted responses for the three-parameter logistic (3PL) model using marginal maximum likelihood estimation via the *mirt* package in R (Chalmers, 2012). Results are compared with empirical-data-

based classical item statistics and IRT parameter estimates.

### 3.3 Results

Two conditions the study examines are sample size per item (500, 1000, and 2000) and GPT training status (Without Training and With Training). For each condition, reported values

| Statistic  | $N$  | Without Training |      |       |      | With Training |      |       |      |
|------------|------|------------------|------|-------|------|---------------|------|-------|------|
|            |      | Mean             | SD   | Min   | Max  | Mean          | SD   | Min   | Max  |
| $p$        | 500  | 0.50             | 0.19 | 0.15  | 0.86 | 0.51          | 0.20 | 0.09  | 0.93 |
|            | 1000 | 0.51             | 0.18 | 0.14  | 0.88 | 0.53          | 0.19 | 0.17  | 0.91 |
|            | 2000 | 0.53             | 0.18 | 0.12  | 0.86 | 0.51          | 0.20 | 0.05  | 0.92 |
| $r_{pbis}$ | 500  | 0.44             | 0.05 | 0.24  | 0.57 | 0.47          | 0.08 | 0.25  | 0.63 |
|            | 1000 | 0.45             | 0.06 | 0.24  | 0.62 | 0.46          | 0.07 | 0.26  | 0.63 |
|            | 2000 | 0.44             | 0.06 | 0.28  | 0.57 | 0.44          | 0.08 | 0.16  | 0.62 |
| $a$        | 500  | 1.24             | 0.24 | 0.69  | 2.12 | 1.45          | 0.43 | 0.74  | 2.75 |
|            | 1000 | 1.31             | 0.32 | 0.56  | 2.51 | 1.47          | 0.48 | 0.66  | 3.02 |
|            | 2000 | 1.30             | 0.28 | 0.75  | 2.17 | 1.34          | 0.41 | 0.82  | 2.51 |
| $b$        | 500  | 0.02             | 0.91 | -2.10 | 2.24 | -0.07         | 0.94 | -2.98 | 2.14 |
|            | 1000 | -0.05            | 0.87 | -2.64 | 1.82 | -0.15         | 0.89 | -2.57 | 1.76 |
|            | 2000 | -0.11            | 0.86 | -2.19 | 1.65 | -0.05         | 1.00 | -2.37 | 3.51 |
| $c$        | 500  | 0.01             | 0.01 | 0.00  | 0.07 | 0.00          | 0.01 | 0.00  | 0.06 |
|            | 1000 | 0.01             | 0.02 | 0.00  | 0.17 | 0.02          | 0.03 | 0.00  | 0.16 |
|            | 2000 | 0.02             | 0.04 | 0.00  | 0.21 | 0.01          | 0.01 | 0.00  | 0.06 |

Table 1: Descriptive Statistics

represent means across 10 simulation runs. Table 1 summarizes proportion correct ( $p$ ), point-biserial correlation ( $r_{pbis}$ ), and IRT  $a$ ,  $b$ , and  $c$  parameter estimates based on the GPT-generated data. The figure shows that the means and variability of  $p$ ,  $a$ , and  $b$  from GPT-generated responses generally seem to be reasonable. For example, at  $N = 1,000$  without training, the means ( $SD$ s) of  $p$ ,  $a$ , and  $b$  are 0.51 (0.18), 1.31 (0.32), and -0.05 (0.87), respectively. In contrast,  $r_{pbis}$  and  $c$  show relatively small means and variability. For example, under the same condition, the mean ( $SD$ ) of  $r_{pbis}$  and  $c$  are 0.45 (0.06) and 0.01 (0.02), respectively. In general, there are minor variations across sample sizes and GPT training status. The only exception to this result is for  $a$ . For example, at  $N = 1,000$ , the mean ( $SD$ ) of  $a$  increases from 1.31 (0.32) without training to 1.47 (0.48) with training.

GPT-based item statistics and parameter estimates are evaluated via the Pearson correlation, bias, root mean squared difference (RMSD), and

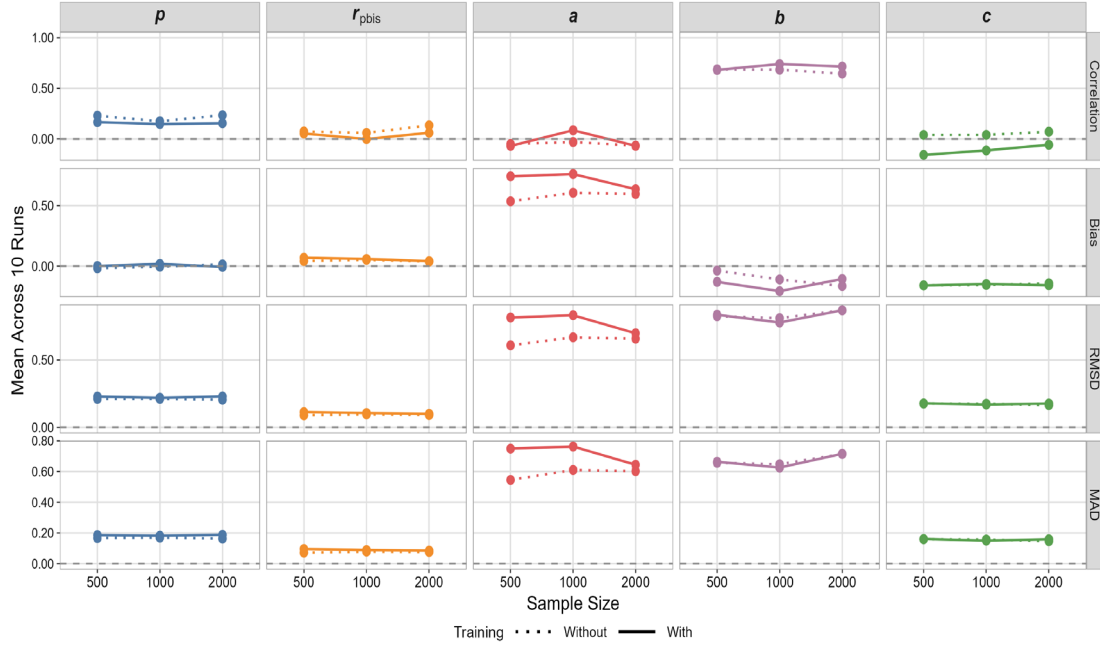


Figure 1: Measures for Classical Item Statistics and IRT Parameter Estimates Without and With Training

mean absolute difference (MAD) using the empirical data-based item statistics and parameter estimates as the benchmark.

Figure 1 presents these measures for the classical item statistics and the IRT item parameter estimates. Increasing sample size and training GPT had minimal impact on performance. Correlations for  $r_{pbis}$ ,  $a$ , and  $c$  are very weak, indicating limited agreement between GPT-generated and empirical estimates. In addition, discrimination parameters ( $r_{pbis}$  and  $a$ ) show positive bias, suggesting that GPT-generated data tend to inflate the degree to which items differentiate examinees by ability.

In contrast, correlations for item difficulty ( $p$  and  $b$ ) are higher, with means of at least 0.15 and 0.65, respectively. For example, the correlation for  $b$  ranges from 0.46 to 0.86, and its association tends to become slightly stronger with training, as shown in Figure 2. Estimates for  $b$  show a slight negative bias, indicating that items appear slightly easier in GPT-generated data.

Computation time (in seconds) is illustrated in Figure 3, which tends to increase with sample size. For example, in the without training condition, the mean computation time increases from 302 seconds ( $SD = 67$ ) at  $N = 500$  to 477 seconds ( $SD = 156$ ) at  $N = 2,000$ . No systematic patterns were observed across the different training conditions.

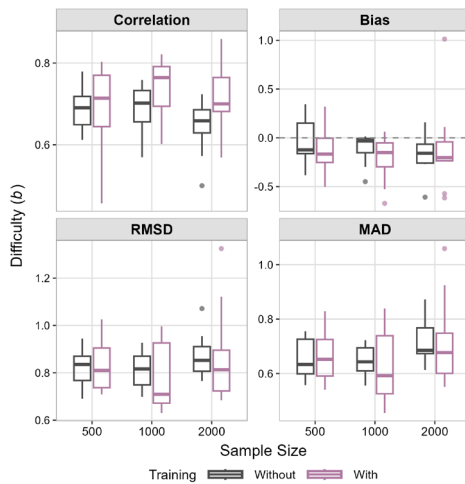


Figure 2: Measures for Item Difficulty Parameter Estimates Without and With Training

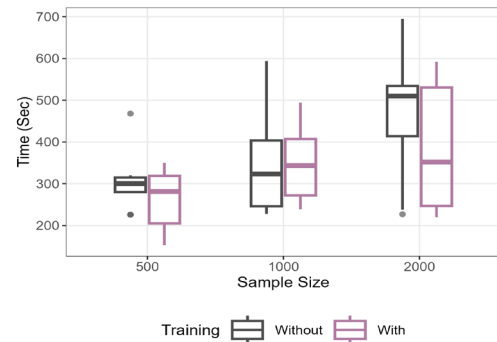


Figure 3: Computation Time Across 10 Runs

## 4 Discussions and Conclusion

This study provides evidence that Skills Insight-based prompting can be used with generative AI to simulate item responses and approximate some psychometric characteristics of reading items. Across conditions, the strongest agreement was observed for item difficulty, particularly for IRT difficulty ( $b$ ) parameter estimates, suggesting that AI-generated response data can meaningfully differentiate easier from more difficult items when grounded in operationally defined score-band descriptions. In contrast, recovery of discrimination-related estimates ( $r_{pbis}$  and  $a$ ) and the  $c$  parameter was weak. Taken together, the findings suggest that the current approach is more useful for preliminary item-difficulty evaluation than for recovering the full set of psychometric item parameters.

Two findings were unexpected: the with training condition did not consistently outperform the without training condition, and increasing simulated sample size did not materially improve agreement with empirical estimates. These results should be interpreted in light of several limitations. First, as noted earlier, “training” in this study refers to task-specific refinement rather than parameter-level fine-tuning of the underlying LLM. In the with training condition, empirical IRT information from the training items was used to compare content-based predicted probabilities with IRT-based probabilities and to refine the model’s prediction process when discrepancies exceeded a prespecified threshold. The model weights themselves were not updated. Thus, the with training condition is better interpreted as an iterative calibration or feedback-based refinement procedure, and the lack of consistent improvement should not be taken as evidence that formal model fine-tuning would be ineffective. Second, the study examined only GPT-5.4 Thinking, so the findings may not generalize to other models or model versions. Third, the simulation was limited to one reading assessment and one fixed form, which constrains generalizability across content areas, test designs, and examinee populations. Future research could replicate the study across models and assessments, and investigate prompting or calibration approaches that improve recovery of the  $a$  and  $c$  parameters.

Overall, the findings from this study highlight the potential of combining content-based student skill profiles with LLMs to support early-stage

item evaluation. By leveraging existing Skills Insight frameworks, the proposed approach offers a practical and interpretable method for obtaining preliminary indications of item difficulty prior to field testing. Although additional research is needed to strengthen the prediction of item parameters, the overall results demonstrate a promising proof of concept and contribute to the growing body of research on AI-assisted assessment development. As generative AI capabilities continue to advance, approaches such as this may become increasingly valuable as a complement to traditional psychometric methods and item development practices.

## References

- Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., & Buttery, P. (2024). Using LLMs to simulate students’ responses to exam questions. Findings of the Association for Computational Linguistics: EMNLP 2024, 11351–11368. <https://doi.org/10.18653/v1/2024.findings-emnlp.663>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Demirkaya, O., Wei, H.-R., & Johnson, E. (2025). Simulating rating scale responses with LLMs for early-stage item evaluation. Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, 385–392.
- Maeda, H. (2025). Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement*, 85(2), 221–244. <https://doi.org/10.1177/00131644241281053>
- OpenAI. (2026). ChatGPT (GPT-5.4 Thinking) [Large language model]. <https://chat.openai.com>
- Peters, S., Zhang, N., Jiao, H., Li, M., & Zhou, T. (2025). Review of text-based approaches to item difficulty modeling in large-scale assessments. Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, 37–47.
- Tang, X., Lu, Y., Behrens, J. T., & Cheng, Y. (2026). Can large language models emulate human performance on educational assessments? PsyArXiv. [https://doi.org/10.31234/osf.io/btja8\\_v2](https://doi.org/10.31234/osf.io/btja8_v2)

## Appendix A. Prompt for Generating Item Responses and Rationale

Project: Predicted Item Response based on Item Content and Skills Insights only

Use item content, item content constraints in the uploaded file [Constraint File.xlsx], and the knowledge, skills and abilities (KSAs) as described in Skills Insight statements only to predict item responses for  $n=[500/1000/2000]$  students. For item matching across files, use item code in the [Item File.json] file to match with item ID in the file [Constraint File.xlsx].

Percentage of students in each score band of the Skills Insights follows the percentages in the Reading excel file uploaded to this Custom GPT. For the simulated students in the same score band, varying the predicted responses considering different extent of mastering KSAs because of different scale scores (e.g., a student with scale score of 239 would be expected to master KSAs slightly better than a student with scale score of 231). No scale score to theta mapping is needed. The prediction is based on whether students with different KSAs are more likely to answer each item correctly or incorrectly based on KSAs described in the Skills Insight statements for each score band and the item content/content constraints. No IRT model is involved.

Include the response matrix, the simulated students' scale scores, the features each item measure, and the rationale of why each simulated student likely answered each item correctly or incorrectly. Produce the output as a downloadable excel file.

## Appendix B. Sample Output of Predicted Item Responses and Rationale

| StudentID | ItemCode | IsCorrect | Rationale                                                                                                                                                                                                                                                                                                                                                                                            |
|-----------|----------|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S0001     | Item XX  | 0         | Likely incorrect because the item depends on using context to determine academic vocabulary, which is above the student's likely demonstrated profile; the student's scale score 206 in band 200–220 indicates beginning foundational reading skills; often struggles unless information is very explicit. Simulated $P(\text{correct})=0.020$ .                                                     |
| S0984     | Item XX  | 1         | Likely correct because the item depends on using context to determine academic vocabulary, which is near the edge of the student's likely demonstrated profile; the student's scale score 290 in band 271–300 indicates can draw subtle or complex inferences, handle relatively uncommon academic vocabulary, and make subtle/complex cross-text connections. Simulated $P(\text{correct})=0.559$ . |