

Using AI-Simulated Response Probabilities for Item Difficulty Estimation

Brad Bolender
Finetune by Prometric

Sara Vispoel
Finetune by Prometric

Abstract

This paper evaluates an AI-based simulated administration approach for estimating item difficulty before empirical calibration. AI-generated option probabilities for synthetic examinees were converted into item difficulty estimates and compared with operational Rasch difficulty statistics. Results showed meaningful alignment, highlighting simulated administration as a potential tool for earlier item performance evidence.

1 Introduction

Pretesting is central to high-quality assessment development. Professional testing standards emphasize the need for evidence supporting the intended interpretation and use of test scores, including evidence related to item functioning, reliability, fairness, and alignment with the intended construct (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In practice, this evidence often depends on administering new items to examinees and evaluating item difficulty, discrimination, distractor functioning, and other empirical item statistics before operational use (Lane, Raymond, & Haladyna, 2016). However, pretesting creates practical constraints. Sufficient response data must be collected before stable item statistics can be estimated, and item quality decisions may not be available until late in the development cycle. These constraints are especially relevant in certification and credentialing contexts, where item exposure, test security, candidate burden, and timely item bank replenishment are ongoing concerns.

Prior work has explored whether item difficulty can be estimated before empirical calibration using machine learning models informed by item features (AlKhuyaey et al., 2024). In this line of

research, difficulty prediction is often framed as a supervised modeling task in which textual, linguistic, semantic, syntactic, or content-based features are extracted from items and used to predict empirical difficulty values. These approaches have shown promise, but they may depend on predictors that are item-type specific, domain-specific, or labor-intensive to extract, limiting generalizability across assessment contexts.

The present study builds on that line of work while shifting the task from feature-based prediction to simulated response modeling. In an earlier study, presented at AIME-Con 2025, we evaluated AI-based difficulty estimation using 9th-grade vocabulary items, where predictions were supported by extracted word-level and lexical information. Here, we extend that approach to conventionally developed certification items by instructing the AI model to evaluate the concepts targeted by each item, the keyed response, and the inferred plausibility of each distractor for examinees at different ability levels. This approach reduces the need for item-type-specific feature engineering and treats difficulty estimation as a simulation of examinee response behavior.

Recent advances in large language models and generative AI have created new possibilities for assessment development and evaluation, while also raising important questions about validity, reliability, fairness, transparency, and test security (Bolender et al., 2023; Hao et al., 2024). In the present context, one potential use is to generate early evidence about item performance by simulating how examinees at different ability levels might respond to multiple-choice items. Such evidence would not replace empirical calibration. Rather, AI-generated response simulations could help identify potentially problematic items earlier, support item review, inform pretest design, or provide weak prior information for later calibration.

In this study, we evaluate an AI-based simulated administration approach for estimating item difficulty before empirical calibration. The approach uses AI to generate option-level response probabilities for synthetic examinees with known ability estimates. These probabilities are used to construct item-level probability curves, from which AI-derived difficulty estimates are extracted. Using item content and operational item statistics from a CompTIA certification exam, we compare AI-derived difficulty estimates with CompTIA's operational Rasch difficulty values. Results show moderate alignment between the two, suggesting that simulated administration can recover meaningful item difficulty signal. We discuss implications for earlier item performance evidence, pretesting efficiency, and future research on using AI-derived estimates as prior information in item calibration.

2 Simulated Administration Approach

The simulated administration approach is intended to generate early item performance evidence before, or alongside, empirical pretesting. Rather than asking an AI system to directly judge whether an item is easy or difficult, the approach asks the system to estimate how examinees at different ability levels would respond to each option in a multiple-choice item. This reframes the item difficulty estimation task as a simulated response modeling problem.

For each item, the AI system is provided with the item stem, response options, keyed response, learning objective information, and calibration examples from similar assessment content. The system is also provided with information about a synthetic examinee, including an estimated ability value. The model then estimates the probability that the examinee would select each response option. The output is not a single predicted response, but a probability distribution across the available options.

This option-level probability structure is important because multiple-choice item performance depends not only on whether the keyed response is identifiable, but also on the plausibility of the distractors. A plausible distractor may attract lower-ability examinees or examinees with partial understanding, while implausible distractors may receive little endorsement across the ability scale. By estimating probabilities for all response options, simulated administration can

produce response patterns that more closely resemble the information available from empirical response data.

The procedure is repeated across synthetic examinees spanning a range of ability levels. For a given item, this produces an estimated response function: the predicted probability of selecting each option as a function of examinee ability. The probability assigned to the keyed response can be treated as an AI-derived probability-correct curve. Fitting a logistic item response curve to that function provides an item-level difficulty estimate derived from the simulated administration. In this study, that AI-derived difficulty estimate is evaluated against the operational Rasch difficulty statistic used in test administration.

Simulated administration is not intended to replace empirical calibration. Empirical response data remain necessary for operational decisions about item quality, scoring, and test form construction. Instead, simulated administration is positioned as a supplemental source of early evidence. If AI-generated response probabilities align with empirical item difficulty, they may support earlier identification of items that are likely to be too easy, too difficult, or otherwise unlikely to perform as intended. They may also inform pretest planning or future calibration models that incorporate AI-derived difficulty estimates as weak prior information.

3 Method

3.1 Data and Items

The study used item content and empirical item statistics from a CompTIA IT certification exam. CompTIA provided item content, learning objective classifications, and operational item statistics from a broader item pool that included multiple item formats. The primary empirical criterion was the Rasch difficulty statistic used by CompTIA in operational item analysis. Empirical item p-values were also used as a secondary check because they provide a directly interpretable indicator of item difficulty.

The present analysis was restricted to single-response, four-option multiple-choice items because the simulated administration procedure modeled a probability distribution across four response options with one keyed response. Items with other formats, such as multiple-select items, were not included in the present analysis. A subset

of 60 single-response, four-option multiple-choice items was selected for the simulated administration experiment. Items were sampled to represent a range of learning objective areas and empirical difficulty levels, ensuring that the experiment included items varying in content and difficulty.

For each analyzed item, the simulated administration procedure used the stem, response options, keyed response, and learning objective information as input to the AI model. Empirical item statistics, including Rasch difficulty and p-value, were reserved for evaluating the AI-derived difficulty estimates and were not treated as target outputs during simulated administration.

3.2 Synthetic Examinees

The simulated administration used 500 synthetic examinees. Synthetic examinees were generated from the empirical examinee dataset to preserve realistic relationships among background characteristics and ability. The available background variables included age group, level of experience, and level of education. These variables were used to model observed person ability estimates in the empirical data.

First, a linear model was fit predicting person theta from examinee background characteristics. This model provided a background-conditional expected theta value for each observed examinee profile. The purpose of the model was not to fully explain examinee ability, but to preserve broad relationships between background characteristics and ability when generating synthetic examinees.

Next, synthetic examinee profiles were created by sampling observed combinations of age, experience level, and education level from the empirical data. Sampling from observed profiles preserved realistic combinations of background characteristics and approximately maintained the proportions present in the original examinee population. For each sampled profile, the fitted model was used to assign a baseline theta value.

To preserve individual variation not explained by the background variables, residuals from the model were sampled and added to the baseline theta values. A small amount of random jitter was then added to avoid exact duplication among synthetic examinees with repeated sampled profiles. This procedure produced synthetic examinees whose theta values reflected both background-related differences and residual person-level variation.

The final synthetic examinee sample was stratified across theta bands to provide coverage from low to high ability levels. This distribution was important because the goal was to estimate response probabilities across the ability scale. Each selected item was paired with each synthetic examinee, producing a complete item-by-examinee design matrix with 30,000 simulated administration records.

3.3 AI Response Probability Generation

For each item-by-examinee combination, an AI model was prompted to estimate the probability that the synthetic examinee would select each response option. The prompt included the item stem, response options, keyed response, learning objective information, and the synthetic examinee's theta estimate. The prompt also included calibration examples from the same assessment context. These examples provided information about how item content, empirical difficulty, option attractiveness, and examinee ability relate to response probabilities.

The AI model was instructed to return a probability distribution across all four response options rather than a single predicted response. This produced four probabilities for each item-by-examinee combination: the estimated probability of selecting option A, B, C, or D. Probabilities were constrained to fall between 0 and 1 and to sum to 1. The probability assigned to the keyed response was retained as the AI-derived probability correct for that item and examinee.

3.4 Extracting AI Difficulty Estimates

For each item, the AI-derived probability correct values were modeled as a function of synthetic examinee theta. A two-parameter logistic curve was fit to the probability-correct function:

$$P_i(\theta) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]}$$

where a represents the slope of the curve and b represents the location or difficulty parameter. Because the observed values were AI-generated probabilities rather than binary scored responses, the curve was fit by minimizing a Bernoulli cross-entropy objective using the AI-derived probability correct as the expected response. The resulting location parameter from this probability-based

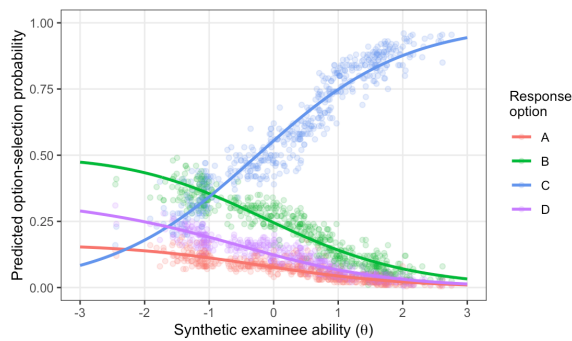


Figure 1. Example option probability curves from the simulated administration procedure. Points are AI-generated option-selection probabilities for 500 synthetic examinees, and lines represent fitted option trace curves.

curve-fitting procedure was retained as the AI-derived difficulty estimate.

Although the curve-fitting procedure also produces a slope parameter, the current study focuses on the difficulty estimate. This decision reflects the purpose of the experiment and the stronger theoretical connection between simulated response probabilities and item difficulty. Discrimination estimates derived from simulated probabilities were not used as primary outcomes.

3.5 Analytic Comparisons

The primary analysis examined the relationship between the AI-derived difficulty estimate and the operational Rasch difficulty statistic provided by the testing program. Pearson correlations were used to evaluate the extent to which the AI-derived estimates recovered the relative ordering of empirical item difficulty.

As a secondary check, AI-derived difficulty estimates were compared with empirical item p-values. Because higher p-values indicate easier items, whereas higher difficulty values indicate more difficult items, a negative correlation between AI-derived difficulty estimates and p-values was expected. This comparison provided an additional validity check on whether the AI-derived estimates reflected item difficulty in the expected direction.

Finally, example option probability curves were inspected to evaluate whether the simulated administration produced plausible response patterns across theta. These curves display the predicted probability of selecting each response option as a function of synthetic examinee ability. For well-behaved items, the keyed option would

generally be expected to increase as theta increases, while distractor probabilities would generally decline or show localized attractiveness among lower- or middle-ability examinees.

4 Results

The simulated administration produced response option probabilities for all 30,000 item-by-examinee combinations. All generated probability distributions were valid, with option probabilities constrained between 0 and 1 and summing to 1 across the four response options. For each item, the probability assigned to the keyed response was modeled across synthetic examinee theta values to extract an AI-derived difficulty estimate.

The primary analysis compared AI-derived difficulty estimates with CompTIA's operational Rasch difficulty statistic. The estimates showed moderate positive alignment with Rasch difficulty, $r = .504$. This indicates that items estimated as more difficult by the simulated administration approach tended to be more difficult according to the operational Rasch criterion.

As a secondary check, AI-derived difficulty estimates were compared with empirical item p-values. As expected, the relationship was negative, $r = -.510$. Items with higher AI-derived difficulty estimates tended to have lower p-values, indicating that they were answered correctly by a smaller proportion of examinees. This result provides additional evidence that the simulated administration approach recovered meaningful item difficulty signal.

Option probability curves were also inspected for individual items. These curves showed the predicted probability of selecting each response option as a function of synthetic examinee theta. Across all analyzed items, the fitted keyed-response probability curve increased as theta increased. Distractor probabilities generally declined or showed localized attractiveness at lower or middle theta values. These patterns are consistent with expected multiple-choice item behavior and suggest that the approach captured more than a single overall difficulty judgment. It also produced interpretable option-level response patterns.

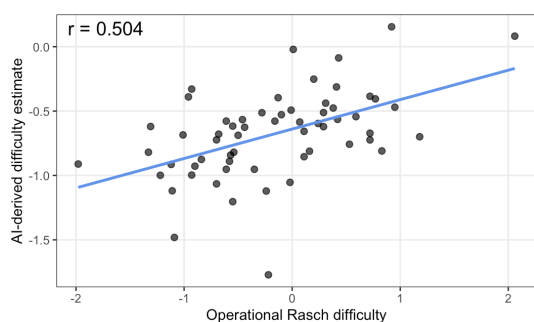


Figure 2. Relationship between operational Rasch difficulty and AI-derived item difficulty estimates for the 60 certification items in the simulated administration experiment. Each point represents one item, and the line shows the fitted linear trend. AI-derived difficulty estimates showed moderate alignment with operational Rasch difficulty, $r = .504$.

5 Discussion

This study evaluated whether AI-simulated response probabilities could recover empirical item difficulty signal for CompTIA items. The results were encouraging. AI-derived difficulty estimates showed moderate alignment with CompTIA's operational Rasch difficulty statistic, and the relationship with empirical p-values was in the expected direction. Items estimated as more difficult by the simulated administration procedure tended to have higher Rasch difficulty values and lower p-values.

These findings suggest that simulated administration may provide useful early evidence about item performance. The procedure did not ask the AI model to directly rate item difficulty. Instead, the model estimated option-level response probabilities for synthetic examinees across a range of ability levels, and item difficulty was derived from the relationship between examinee theta and the predicted probability of selecting the keyed response. This distinction is important because the approach attempts to approximate response behavior rather than rely on a direct qualitative judgment of item difficulty.

The option-level nature of the output is also important. Multiple-choice item performance depends not only on whether the keyed response is recognizable, but also on how plausible the distractors are for examinees at different ability levels. By producing probabilities for all response options, simulated administration can generate interpretable option probability curves. These

curves may help identify items with weak distractors, overly attractive distractors, or response patterns that do not vary as expected across ability levels.

The results should not be interpreted as evidence that AI-simulated administration can replace empirical pretesting or operational calibration. Empirical response data remain necessary for item calibration, validity evidence, test assembly, and operational decision-making. However, the observed alignment with operational difficulty statistics suggests that simulated administration may serve as a supplemental source of evidence earlier in the assessment development process. Potential applications include identifying items likely to be too easy or too difficult, prioritizing items for review, informing pretest design, and supporting decisions about which unscored items should receive additional empirical evaluation.

A promising future direction is to investigate whether AI-derived difficulty estimates can be incorporated as weak prior information during item calibration. In this framing, AI estimates would provide an initial expectation about item difficulty that could be updated as empirical response data are collected. This approach may be especially useful when pretest sample sizes are limited. Future research could compare calibration accuracy across different sample sizes, with and without AI-informed priors, to determine whether stable item calibration can be achieved with fewer examinee responses.

Overall, the findings support simulated administration as a promising method for generating early item performance evidence. The approach recovered meaningful difficulty signal relative to CompTIA's operational Rasch difficulty statistic and produced interpretable option-level response patterns. These results provide a foundation for future work examining whether AI-derived response evidence can reduce pretesting burden, improve item review, and support more efficient calibration workflows.

6 Limitations and Next Steps

Several limitations should be considered when interpreting these findings. The study used a selected subset of 60 items from the CompTIA certification exam. Although the items were sampled to represent a range of learning objective areas and difficulty levels, additional work is needed to determine whether the results generalize

across larger item pools, other certification exams, other item formats, and different content domains.

The synthetic examinees were generated from observed examinee background characteristics and ability estimates, but they remain simulated profiles rather than real candidates. The generation procedure was designed to preserve realistic relationships among age, experience, education, and theta; however, simulated examinees cannot fully capture the range of knowledge, misconceptions, test-taking behavior, and response strategies present in actual examinee populations.

AI-generated response probabilities may also be sensitive to implementation choices, including prompt structure, calibration examples, learning objective information, and model version. Future studies should evaluate the stability of results across prompt variants, exemplar sets, and AI models.

The current study focused primarily on item difficulty. Although the simulated administration procedure produced option-level response probabilities, and although probability curves can be used to estimate slope-like parameters, the present evidence is strongest for recovering relative item difficulty. Future work should examine whether simulated administration can provide useful evidence about discrimination, distractor functioning, item-total relationships, or other indicators of item quality.

The observed alignment with operational Rasch difficulty was promising but not sufficient to support replacing empirical pretesting or calibration. Empirical response data remain necessary for operational item decisions. The appropriate interpretation is that simulated administration may provide supplemental early evidence that can inform item review, pretest planning, or future calibration studies.

The next phase of this work will examine whether AI-derived difficulty estimates can be used as weak prior information during item calibration. In this approach, AI estimates would serve as an initial expectation about item difficulty, while empirical response data would continue to determine final operational parameters. A simulation study could compare calibration results across different real-examinee sample sizes, with and without AI-informed priors, to determine whether stable item calibration can be achieved with fewer examinee responses.

Additional next steps include evaluating the method on a larger item set, testing the approach on additional certification exams or content domains, and investigating whether option-level probability curves can help identify potentially problematic unscored items. For example, simulated administration may help flag items that are likely to be too easy, too difficult, weakly related to ability, or characterized by unusual distractor behavior. These applications would position AI-simulated administration as an early evidence tool within the broader item development and validation process.

Acknowledgments

The authors thank Russell Smith, Carl Bowman, Danielle Castelli, and CompTIA for providing access to item content, operational item statistics, and related assessment context, and for their thoughtful discussion, feedback, and review. The authors are grateful for the opportunity to collaborate with this group and appreciate the measurement and assessment expertise they brought to the project.

References

- AlKhuzaey, S., Grasso, F., Payne, T. R., & 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.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.
- Bolender, B., Foster, C., & Vispoel, S. (2023). The criticality of implementing principled design when using AI technologies in test development. *Language Assessment Quarterly*, 20(4–5), 512–519.
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29.
- Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (2nd ed.). Routledge.