Using Generative AI to Develop a Common Metric in Item Response Theory

Peter Baldwin

Office of Research Strategy, National Board of Medical Examiners, Philadelphia, USA
pbaldwin@nbme.org

Abstract

Item response theory (IRT) models are subject to scale indeterminacy, causing parameters to be arbitrarily scaled. Consequently, parameters from independently calibrated test forms are not directly comparable without first estimating the linear transformation that aligns their respective scales. This paper introduces a novel procedure that uses large language models (LLMs) to estimate the transformation's slope and intercept. The method is evaluated using empirical data from a medical licensure exam. Results indicate that the LLM-based approach consistently recovers the slope across conditions, while intercept recovery is moderately sensitive to differences in average item difficulty between forms and improves as that difference narrows.

1 Introduction

When examinees take different test forms designed to measure the same trait, their scores must be adjusted for comparability. For number-correct scores (or transformations thereof), this process is called *equating*. In IRT, parameters are invariant up to a linear transformation; thus, while equating per se is unnecessary, a similar scaling adjustment is still required to ensure comparability across independently calibrated forms. After this adjustment, model parameters are expressed on a common scale—sometimes called developing a *common metric* (Stocking and Lord, 1983).

This scaling is necessary because the origin and unit of the latent scale must be arbitrarily fixed—directly or indirectly—to identify the model. Independently calibrated forms will therefore generally differ in scale, requiring a linear transformation before parameter estimates can be compared. This paper addresses this problem and proposes a procedure that uses GPT-based LLMs to estimate the slope and intercept of the required transformation. The method is illustrated using empirical data from the medical licensure domain.

2 Background

2.1 Problem Definition

Although scale indeterminacy affects all IRT models, we illustrate the issue using the two-parameter logistic model (2PL):

$$P(\theta) = \frac{1}{1+e^{-Da(\theta-b)}}, \qquad (1)$$

where $\theta \in \mathbb{R}$ denotes proficiency, $a \in \mathbb{R}_{>0}$ is item discrimination (equal to 4 times the item response function's (IRF) maximum slope), and $b \in \mathbb{R}$ is the item difficulty, the point on the difficulty/proficiency scale where the IRF inflects). $P(\theta)$ gives the probability of a correct response for an examinee with proficiency θ.[1]

A key feature of IRT is parameter invariance: item parameters are independent of examinee sample, and examinee proficiencies are independent of item set (Hambleton et al. 1991). However, this invariance holds only up to a linear

[1] Note: D is a scaling factor equal to $D = 1.702$ that allows the more mathematically tractable logistic formulation to closely resemble the normal ogive function, which preceded the logistic function in the historical development of IRT (Birnbaum, 1968).

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 281–289
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

transformation: for any slope j and intercept k, the transformation $a' = a / j$, $b' = bj + k$, and $\theta' = \theta j + k$ leaves $P(\theta)$ unchanged.

This scale indeterminacy is expected—these model parameters are not directly observable, requiring arbitrary scaling—but it complicates comparisons across independently calibrated test forms. Conventions exist for identifying IRT models (e.g., setting θ's mean and SD to 0 and 1, respectively) but they do not guarantee a shared scale across forms, since these constraints are applied separately to each. A linear transformation is still required. We denote its slope and intercept γ and η, respectively.

To estimate γ and η, something common across forms is needed (Baldwin and Clauser 2022), typically in the form of shared parameters (e.g., anchor items). These common parameters, being invariant up to a linear transformation, can be used to estimate the linear relationship between scales. Many well-known linking methods take this approach (Hambleton and Swaminathan 1985; Kolen and Brennan 2014). Absent common items, ancillary covariates that correlate with model parameters can sometimes be used (e.g., Mislevy et al. 1993; Wiberg and Bränberg 2015).

A single-group design, in which both forms are administered to the same examinees, allows estimation of the transformation constants via shared proficiencies. However, this approach is often infeasible due to examinee burden. More common is the *non-equivalent groups with anchor test* design. Although less demanding for examinees, it relies on item parameter invariance—an assumption that may be violated due to item exposure, evolving curricula, or changes in exam preparation, leading to *item parameter drift*.

To address these limitations, we propose using generative AI to create shared parameters across forms. Specifically, GPT-based LLMs are tasked with estimating item-level success probabilities for typical examinees from defined groups. These probabilities are used to derive a common set of synthetic proficiency parameters across forms—analogous to a single-group design—enabling estimation of the transformation constants without requiring common items, examinees, or external covariates.

The proposed approach is illustrated using empirical data from the medical licensure domain. It performed well, particularly for slope estimation, with high consistency across all conditions. Intercept estimates were more sensitive to differences in average item difficulty between forms, with smaller gaps yielding more accurate results.

2.2 Related Work

A review of the literature did not identify any studies that use LLMs directly to develop a common metric. However, several studies address related challenges, particularly item difficulty prediction—a long-standing topic in educational and psychological measurement (e.g., Beinborn et al., 2015; Huang et al., 2017; Ha and Yaneva, 2018). Current LLM-based approaches to difficulty prediction fall into two categories: (a) item-parameter prediction and (b) item-specific examinee-group performance prediction. The latter, while not identical to the task described here, is more closely aligned. Each approach is discussed below.

Item-parameter prediction estimates classical or IRT-based indices from item text. For example, Razavi and Powers (2025) used GPT-based models to predict difficulty for K–5 math and reading items. Their feature-based ensemble models outperformed direct rating methods, reaching correlations up to $r = 0.87$ with empirical difficulties. However, accuracy may decline in domains requiring specialized knowledge or complex reasoning. For instance, in a shared task using medical multiple-choice questions (MCQs), Yaneva et al. (2024) reported that difficulty estimation remains challenging in this domain.

The second approach—item-specific performance prediction—models how systems or subgroups perform on individual items. Studies have linked item difficulty for question-answering systems to human performance (e.g., Yaneva et al., 2019; Uto et al., 2024; Liu et al., 2025; Maeda, 2025), though not always with high precision. More relevant here are studies modeling interactions between examinee subgroups and items. Feng et al. (2025) used chain-of-thought prompting and synthetic response generation to predict MCQ difficulty for defined cohorts. Park et al. (2024) used AI models as proxies for students at different skill levels. While promising, such methods raise concerns about bias in synthetic responses and highlight the need for further validation.

3 Methodology

3.1 Proposed Procedure

Let $P_{g,m,i}$ denote the predicted probability that a typical examinee from group g will answer item i correctly, according to LLM m. Likewise, for test form f, let $\mathbf{P}_{g,m,f}$ be the vector of predicted probabilities for that set of items. Now, suppose an IRT model is fit to an empirical dataset for form f, yielding item parameter estimates. These estimates can then be used to estimate a proficiency value $\hat{\theta}_{g,m,f}$ that, in some way, best describes $\mathbf{P}_{g,m,f}$.

Because this process can be replicated multiple times, let $\hat{\theta}_{g,m,f,r}$ denote the estimate of $\theta_{g,m,f}$ associated with the rth such replication. To improve stability, we can then take the average across R replications:

$$\bar{\hat{\theta}}_{g,m,f} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{g,m,f,r} . \quad (2)$$

This average reduces the impact of sampling variability from LLM output and estimation noise.

Because IRT proficiencies are form-invariant, were $\theta_{g,m,f}$ to be estimated using two different test forms, the resulting two $\bar{\hat{\theta}}_{g,m,f}$ will be the same (excepting random error) *up to a linear transformation*:

$$\bar{\hat{\theta}}_{g,m,f_B} \approx \bar{\hat{\theta}}_{g,m,f_N}\gamma + \eta , \quad (3)$$

where the subscripts f_B and f_N, indicate base form or new form, respectively, and γ and η represent the slope and intercept of the transformation needed to place f_N's estimates on the scale of f_B.

Given G examinee groups and M LLMs, this procedure yields $G \times M$ mean estimated proficiencies, $\bar{\hat{\theta}}_{g,m,f}$, for each form. The transformation constants γ and η can then be estimated using the *mean and sigma* method (Marco 1977; Kolen and Brennan 2014), which matches the means and standard deviations of the two sets:

$$\hat{\gamma} = \frac{SD\left(\bar{\hat{\theta}}_{g,m,f_B}\right)}{SD\left(\bar{\hat{\theta}}_{g,m,f_N}\right)} \quad (4)$$

$$\hat{\eta} = \bar{\hat{\theta}}_{f_B} - \gamma\bar{\hat{\theta}}_{f_N} , \quad (5)$$

where

$$\bar{\hat{\theta}}_{f} = \frac{1}{GM}\sum_{g=1}^{G}\sum_{m=1}^{M}\bar{\hat{\theta}}_{g,m,f} \quad (6)$$

$$SD\left(\bar{\hat{\theta}}_{g,m,f}\right) = \sqrt{\frac{1}{GM}\sum_{g=1}^{G}\sum_{m=1}^{M}\left(\bar{\hat{\theta}}_{g,m,f} - \bar{\hat{\theta}}_{f}\right)^2} . \quad (7)$$

Notably, this application differs from traditional difficulty prediction in that there is no requirement that $\bar{\hat{\theta}}_{g,m,f}$ reflect the actual proficiency of a typical examinee in group g. That is, accuracy of $\bar{\hat{\theta}}_{g,m,f}$ is less important than form-invariance.

4 Experiments

4.1 Examinee Response Data

We evaluated the procedure using empirical data from the Step 2 exam of the United States Medical Licensing Examination (USMLE®) sequence. Step 2 is typically taken by medical students after their third year of medical school, following their core rotations, and consists of multiple simultaneously administered test forms, each with ~318 MCQs. This study used ~220 items (text-based and table-based) from a single form, along with responses from ~1,500 examinees.

Responses were modeled using the 2PL (Equation 1). Because all items came from a single form, item parameters were estimated on a common scale. Parameters were scaled such that proficiencies had a mean of 0 and SD of 1, simplifying interpretation.

4.2 LLM Data

For each MCQ, a prompt was generated instructing the LLM to act as "an expert medical education analyst" with "thorough knowledge of how medical students and residents perform on USMLE®-style multiple-choice questions." The LLM was then told: "You are tasked with predicting the performance of the typical examinee from each of five different examinee groups on the

following USMLE® multiple-choice question." The prompt included the MCQ, its correct answer, exam label (Step 2), item type (e.g., "diagnosis"), and topic area (e.g., "cardio: infectious disorders"). The five examinee groups—first- through fourth-year medical students and first-year medical residents (PGY-1)—were listed, followed by the judgment task: "Think carefully (internally) about each group's level of training, typical preparedness, and likelihood of arriving at the correct answer… Factor in both knowledge and potential guessing... Provide one probability… for each of the five groups… [that] represents the probability that a typical examinee within that group will answer the question correctly."

For each of fifty replications, the prompt was submitted separately to three large language models (LLMs)—GPT-o1, GPT-o3, and GPT-4.1—via the OpenAI API (OpenAI, 2024). To ensure item security, we used private deployments of these models through Azure OpenAI. While this was the implementation used here, the procedure itself is model-agnostic.

In this way, $G=5$, $M=3$, and $R=50$, yields a set $5\times3\times50=750$ $P_{g,m,i}$ for each of the approximately 220 MCQs totaling approximately $5\times3\times50\times220=165{,}000$ predicted probabilities across the ~220 items.

4.3 Experiments

The full set of items was randomly divided into two equal-length artificial test forms: a *base form* and a *new form*. Form difficulty was manipulated as two study factors: (a) across-form difference in mean item difficulty and (b) across-form difference in the standard deviation of item difficulties. Each factor had 11 symmetrically spaced levels (from -0.25 to 0.25 in 0.05 increments), varied independently, resulting in 21 conditions (10 for mean differences, 10 for SD differences, and 1 baseline condition). An exploratory analysis using a fully crossed design found that slope estimates were largely insensitive to changes in mean item difficulty, and intercept estimates were similarly unaffected by changes in item difficulty spread. For this reason, rather than employing a fully crossed design, we explored only conditions in which exactly one parameter was varied at a time, while holding the other parameter fixed at zero.

While no specific proficiency estimation procedure is required, this study used a two-step empirical Bayes approach with Newton–Raphson optimization designed to ensure monotonic deviance reduction through step-size constraints and backtracking. Initial estimates were computed using diffuse (flat) priors. The empirical mean and standard deviation of these estimates then served as Gaussian priors in a refined second phase. At each iteration, values were optimized by minimizing deviance—the sum of the negative log-likelihood of LLM-predicted response probabilities and a Gaussian prior penalty. Newton updates were derived analytically from the 2PL model (Equation 1), using closed-form gradients and Hessians based on item parameters from the empirical dataset. Step sizes were clamped, and backtracking ensured monotonic deviance reduction. The estimation procedure terminated once convergence criteria (step sizes < 1×10^{-6}) were met.

Once $\hat{\theta}_{g,m,f,r}$ values were computed for all replications, they were averaged as described in Equation 2, yielding 15 values for each artificial test form. These proficiency estimates were then used to estimate the slope and intercept of the transformation line using Equations 4 and 5.

Using the same set of 165,000 predicted probabilities, the procedure—item assignment, proficiency estimation, and transformation recovery—was repeated multiple times. Because item assignments were randomized (within specified difficulty constraints), the resulting $\hat{\gamma}$ and $\hat{\eta}$ varied across repetitions. This variation reflects sensitivity to item selection, although it is smaller than would be expected had each repetition drawn from a new item pool. For this reason, the mean $\hat{\gamma}$ and $\hat{\eta}$ across repetitions were calculated for each of the 21 difference-in-form-difficulty conditions, and a new repetition—considered *typical*—was generated that produced $\hat{\gamma}$ and $\hat{\eta}$ within 0.005 of these means. These "typical" artificial test forms served as stable reference points for evaluating sampling variability more accurately via bootstrapping. One thousand bootstrap draws were created by independently sampling, with replacement, both LLM replications and items within each form. For each bootstrap draw, a $\hat{\gamma}_d$ and $\hat{\eta}_d$ were estimated. These bootstrap-estimated transformation constants were then used to approximate the sampling distributions of $\hat{\gamma}$ and $\hat{\eta}$.

4.4 Evaluation Criteria

Because all items were jointly calibrated, the "true" transformation function was the identity function: $\gamma = 1$ and $\eta = 0$. Accuracy was evaluated by the proximity of estimated values to these targets.

5 Results

5.1 Recovery of Transformation Function Slope

Figure 1 shows the estimated slopes as a function of the difference in item-difficulty standard deviations between forms. Also shown are the boundaries for the middle 95% of the distribution

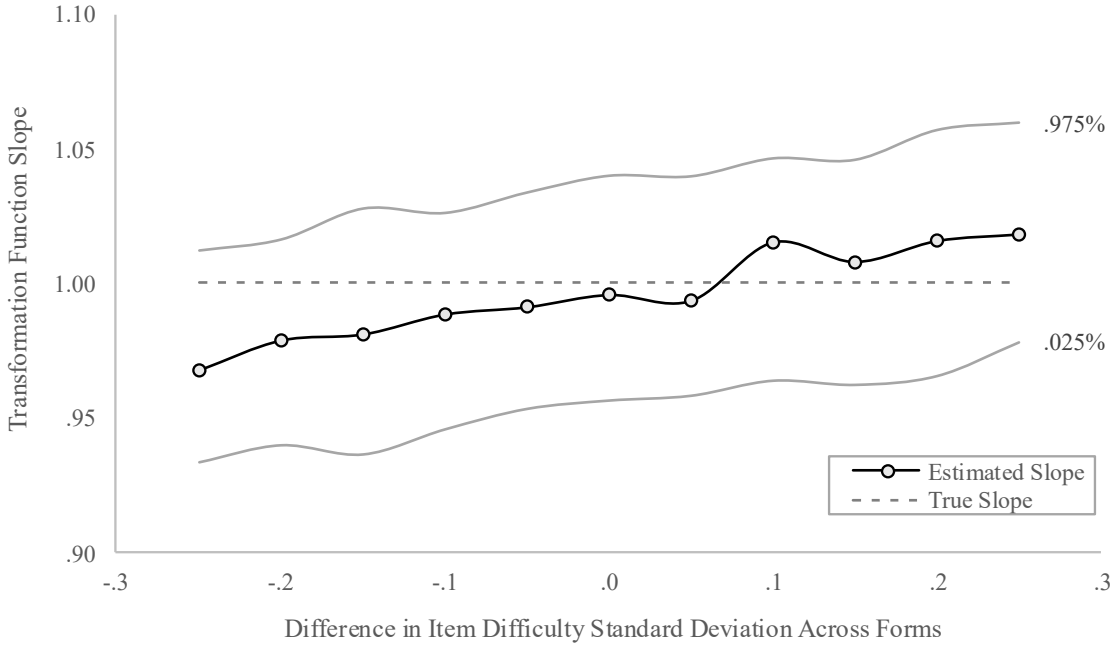

Figure 1: Estimated transformation function slope (black line) as a function of difference in item difficulty standard deviations. The boundaries for the middle 95% of bootstrap slopes are also given (light grey lines); the broken grey line shows the true slope.

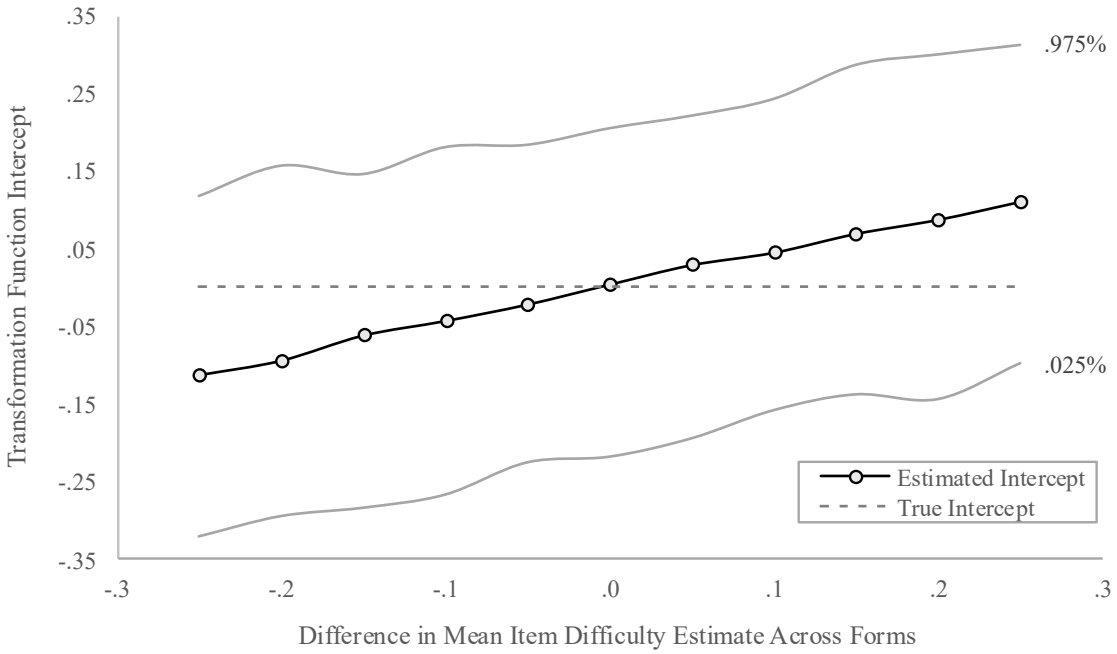

Figure 2: Estimated transformation function intercept (black line) as a function of difference in mean item difficulty. The boundaries for the middle 95% of bootstrap intercepts are also given (light grey lines); the broken grey line shows the true slope. Note: the vertical axis scale spans .70 whereas for Figure 1, this axis spans only .20.

of bootstrap slopes ($\hat{\gamma}_d$; light grey) and the true slope ($\gamma = 1$; broken grey line). Across all conditions, the estimated slope deviates by no more than 0.03 from the true value. The estimates are most accurate when the difference in item-difficulty standard deviations between forms is close to zero. Likewise, the span of the 95% bootstrap sampling distribution is less than 0.09 for all conditions.

5.2 Recovery of Transformation Function Intercept

Figure 2 shows the estimated intercepts as a function of the difference in mean item difficulty between forms, following the same structure as Figure 1. The boundaries for the middle 95% of bootstrap intercepts ($\hat{\eta}_d$; light grey) and the true intercept ($\eta = 0$; broken grey line) are also shown. Intercept estimates improve as the difference in mean item difficulty across forms approaches zero. However, the vertical axis in Figure 2 spans 3.5 times the range of Figure 1, indicating greater variability. In the most extreme condition, the absolute difference between the true and estimated intercepts ($|\hat{\eta} - \eta|$) reaches 0.11. This difference does not fall below 0.05 until the across-form difference in mean item difficulty is $\leq$ 0.10. Similarly, the span associated with the middle 95% of the bootstrap intercepts is greater than that observed for the slopes—with spans up to 0.45.

6 Conclusion

6.1 Discussion

This paper describes a procedure for estimating the transformation constants required to place independently calibrated test forms on a common scale. It follows a single-group (common-person) design (Kolen and Brennan, 2014), but instead of using common examinees, proficiency estimates for a typical test taker from each of five predefined groups are used. These estimates are based on judgment tasks given to three LLMs: GPT-o1, GPT-o3, and GPT-4.1. The method was demonstrated using real examinee-response data from the USMLE® Step 2 exam.

The procedure recovered transformation-function slopes with high precision: across all conditions, slope estimates deviated from true values by no more than 0.03. Intercept estimates were more sensitive to model–data misfit and exhibited greater variability, particularly when mean form difficulties differed substantially. This likely reflects residual dependencies between the LLM-generated proficiency estimates and the item pool, undermining the assumption of conditional independence.

In IRT, proficiency parameters are item-set invariant. While proficiency *estimates* are never fully independent of the items used to derive them, the LLM-generated estimates in this study appeared especially sensitive to item characteristics. This suggests a degree of conditional dependence that may stem from misalignment between LLM-predicted probabilities and the modeled item response function. Because the success of the proposed procedure relies on form-invariant proficiency estimates, such dependencies likely contributed to the observed difficulties in recovering intercepts.

Developing a common metric across test forms administered at different points in time presents a challenge: *common items must exhibit invariance over time*. For testing programs where item parameter drift is a concern, this is a vexing problem. The procedure proposed here does not require items to have this property. Instead, it relies on LLMs to produce form-invariant proficiency estimates, and it is these estimates—rather than common item parameters—that are used to estimate the transformation constants needed to create a common scale across forms. If successful, the proposed procedure represents a considerably more secure method for maintaining a common scale over time. This will be especially attractive to testing programs that administer high-stakes tests following an episodic testing design.

6.2 Limitations

Although the procedure is not specific to any testing program, content domain, IRT model, or LLM, it was demonstrated using medical-domain items from the USMLE®, the 2PL IRT model, and three OpenAI LLMs. These design choices limit the generalizability of the findings.

The USMLE® assesses highly specialized technical content and is taken by a relatively homogeneous examinee population. It is unclear whether the findings extend to more general domains. However, previous studies have reported stronger performance for LLM-based predictions in broader content areas (e.g., Uto et al., 2024; Liu et al., 2025; Maeda, 2025; Razavi and Powers

2025), suggesting that the current results may underestimate the method's effectiveness in less technical contexts.

While the 2PL model is widely used, some testing programs—particularly in K–12 settings—prefer models like the 3PL, which incorporate additional complexity and assumptions about guessing behavior. Although the proposed procedure is not restricted to any single IRT model, it remains unclear whether LLM-based predictions align equally well under models other than the 2PL.

The OpenAI models used in this study are widely known, but they are neither the only nor necessarily the most effective LLMs for this task. Alternative models—used individually or in ensembles—may offer improved accuracy and consistency. Moreover, prediction quality is likely influenced by prompt phrasing and model settings. This study employed a fixed prompt and the default temperature, but future work should examine how variations in prompt structure and sampling parameters affect prediction accuracy and downstream performance. Finally, data security remains a critical concern. This study used private LLM deployments with no data logging or model training from inputs. However, not all models offer this level of protection—an important consideration for testing programs concerned with safeguarding test content.

Finally, the number of examinee groups and the number of items per form were chosen to suit the illustrative nature of this study. These design aspects may influence the quality of estimated transformation constants and the method's scalability. Other programs are likely to involve different group structures or item counts, and the procedure's performance under such conditions remains untested.

6.3 Future Work

Although the procedure performed well under most conditions, it may still fall short of the precision required for high-stakes applications, and several avenues for improvement remain.

First, the procedure is not limited to a fixed number of LLMs. Although this study used three widely known models, incorporating additional models—or employing ensemble strategies—may further improve the quality and stability of the proficiency estimates used to derive the transformation constants.

Similarly, although five examinee groups were used here, additional or alternative groupings could further enhance performance. Exploring optimal group configurations and model-specific strengths across subpopulations may yield more robust results. Increasing the number of examinee groups would also increase the number of independent estimates contributing to the transformation calculation, potentially improving precision.

The greatest limitation of the procedure lies in intercept estimation, specifically the bias in $\hat{\eta}$ when test forms differ in average difficulty. This issue may be mitigated through improved form design. For example, assembling forms to have closely matched mean difficulties can reduce the conditions under which intercept estimation becomes unstable. Importantly, precise individual difficulty estimates are not needed for this purpose; only mean difficulty must be controlled. This may be more tractable using existing difficulty-prediction methods.

Finally, alternative test assembly and delivery strategies could further improve the method's performance. For example, consider a scenario in which a large number of forms—comprising both unique and anchor items—are administered concurrently. These forms could be placed on a common metric using traditional IRT linking techniques (e.g., nonequivalent groups with anchor tests). Now suppose that multiple such administrations occur over time, as in an episodic testing design. In this case, large pools of administration-specific items, already placed on a common scale within each administration, could serve as input to the proposed procedure, yielding substantially larger item sets for estimating the required across-administration transformation. Future research should investigate such strategies, which focus on optimizing test design conditions rather than altering the procedure itself.

Acknowledgments

The author thanks the National Board of Medical Examiners for supporting this work.

References

Peter Baldwin and Brian E. Clauser. 2022. Historical perspectives on score comparability issues raised by innovations in testing. *Journal of Educational Measurement* 59(2):140–160. https://doi.org/10.1111/jedm.12318

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics* 2:517–530. https://doi.org/10.1162/tacl_a_00200

Allan Birnbaum. 1968. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479, Reading, MA. Addison-Wesley. https://ia601405.us.archive.org/32/items/in.ernet.dli.2015.139135/2015.139135.Statistical-Theories-Of-Mental-Test-Scores.pdf

Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew Lan. 2025. Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. *arXiv preprint* arXiv:2503.08551. https://arxiv.org/abs/2503.08551

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 389–398, New Orleans, Louisiana. Association for Computational Linguistics. https://aclanthology.org/W18-0548/

Ronald K. Hambleton and Hariharan Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Boston, MA. Kluwer-Nijhoff. https://link.springer.com/book/10.1007/978-94-017-1988-9

Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA. Sage Publications.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1):1352–1359. https://doi.org/10.1609/aaai.v31i1.10740

Radhika Kapoor, Sang T. Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W. Domingue. 2025. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint* arXiv:2502.20663. https://arxiv.org/abs/2502.20663

Michael J. Kolen and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices* (3rd ed.). New York, NY. Springer. https://link.springer.com/book/10.1007/978-1-4939-0317-7

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* 56(3):1028–1052. https://doi.org/10.1111/bjet.13570

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ. Lawrence Erlbaum Associates. https://www.routledge.com/Applications-of-Item-Response-Theory-To-Practical-Testing-Problems/Lord/p/book/9780898590067

Hotaka Maeda. 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

Gary L. Marco. 1977. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement* 14(2):139–160. https://www.jstor.org/stable/1434012

Robert J. Mislevy, Kathleen M. Sheehan, and Marilyn S. Wingersky. 1993. How to equate tests with little or no data. *Journal of Educational Measurement* 30(1):55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

OpenAI. 2024. OpenAI API models documentation. https://platform.openai.com/docs/models

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.477/

Pooya Razavi and Sonya J. Powers. 2025. Estimating item difficulty using large language models and tree-based machine learning algorithms. *arXiv preprint* arXiv:2504.08804. https://arxiv.org/abs/2504.08804

Martha L. Stocking and Frederic M. Lord. 1983. Developing a common metric in item response theory. *Applied Psychological Measurement* 7(2):201–210. https://doi.org/10.1177/014662168300700208

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2024. Question difficulty prediction based on virtual test-takers and item response theory. In *CEUR Workshop Proceedings*, Vol. 3772. https://ceur-ws.org/Vol-3772/paper1.pdf

Marie Wiberg and Kenny Bränberg. 2015. Kernel equating under the nonequivalent groups with covariates design. *Applied Psychological Measurement* 39(5):349–361. https://doi.org/10.1177/0146621614567939

Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the difficulty of multiple-choice

questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019)*, pages 11–20, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/W19-4402/

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.39/