

Anchored Bradley-Terry Calibration Using LLM Comparative Judgments

Ummugul Bezirhan, Matthias von Davier

Boston College
Chestnut Hill, Massachusetts, USA
{bezirhan, vondavim}@bc.edu

Abstract

AI-based comparative judgment was evaluated as a tool for early item difficulty estimation. Three AI judges compared new TIMSS Grade 4 mathematics items with calibrated anchors, with rankings analyzed using a fixed-anchor Bradley-Terry model. Results show meaningful difficulty signals, supporting scalable supplementary use while highlighting anchor coverage and comparison-network design.

1 Introduction

Assessment developers typically design items to target particular regions of the proficiency continuum, as different ranges of difficulties are intended to measure different knowledge, skills, and abilities. Therefore, obtaining an early indication of item difficulty can inform item development, test assembly, and score interpretation (Attali et al., 2014). For newly developed items, however, conventional calibration methods generally require substantial amounts of student response data that are not yet available. This creates a practical challenge for assessment programs: how can item difficulty be estimated early enough to guide development decisions while maintaining acceptable measurement quality?

Comparative judgment offers a promising alternative. Rather than asking judges to assign an absolute score to an item, comparative judgment asks them to decide which of two items is more difficult. These repeated pairwise decisions can then be modeled statistically to recover a latent scale of item difficulty. The approach has a long history in psychology and educational assessment, where it has been used to produce reliable scales for student work and other complex performances (Thurstone, 1927; Hambleton et al., 2003).

This study examines whether AI-based comparative judgment can support item difficulty measurement. Using three AI judges, newly developed

items were compared against anchor items already situated in the item pool, and the resulting rankings were analyzed using a Bradley-Terry model to estimate the relative difficulty of the new items on a shared metric (Bradley and Terry, 1952; Turner and Firth, 2012).

The study contributes to the literature on AI in measurement in two ways. First, it tests whether AI-generated pairwise judgments can function as a meaningful signal for item difficulty. Second, it assesses whether anchor-based comparative judgment can place new items onto an existing difficulty scale in a way that is useful for assessment development. In doing so, it extends prior work showing that human judges estimate item difficulty far more accurately through comparison than through absolute rating (Attali et al., 2014), asking whether LLM judges can substitute for human ones within an anchored Bradley-Terry framework. By combining LLM judgment with a well-established pairwise comparison model, the study aims to extend comparative judgment methods into a scalable, AI-supported measurement workflow.

2 Background

Obtaining item difficulty estimates typically requires administering items to large student samples and fitting psychometric models to response data. While these procedures provide high-quality parameter estimates, they are resource intensive and can only be completed after data collection. Assessment programs therefore seek supplementary methods that can provide preliminary difficulty information during item development (Attali et al., 2014).

Historically, expert judgment has served as one such source. Subject matter experts review items and judge expected student performance, cognitive complexity, and relative difficulty. Although expert review is an essential component of assessment

development, expert ratings usually produce low and variable correlations with empirical difficulty (e.g., [Bejar, 1983](#); [Mansoor et al., 2024](#)). Comparative judgment provides an alternative to direct item rating. Rather than assigning absolute difficulty values, judges compare pairs of items and determine which item is more difficult.

Pairwise judgments can be aggregated using probabilistic choice models. One of the most widely used approaches is the Bradley-Terry model, which estimates latent item locations from observed pairwise outcomes. In the Bradley-Terry framework, the probability that item i is judged more difficult than item j depends on the difference between their latent difficulty parameters. Repeated comparisons across many item pairs allow the model to recover a latent ordering of items on a common scale.

Prior research has explored the use of LLMs for a variety of assessment-related tasks, including predicting item difficulty. Some studies prompt LLMs to estimate item difficulty directly ([Bezirhan and von Davier, 2025](#)), whereas others infer difficulty indirectly from model-based indicators such as response accuracy, confidence, or question-answering loss (e.g., [Rogoz and Ionescu, 2024](#); [Park et al., 2024](#)). Additional work has combined item-text features with LLM embeddings and contextual information to predict item difficulty ([Kapoor et al., 2025](#)). Although these approaches have produced promising results, findings remain mixed, particularly for estimating operational item difficulty. Item difficulty reflects not only characteristics of the item itself but also interactions between item features and the target student population ([AlKhuyaey et al., 2024](#)).

One promising direction involves using LLMs as comparative rather than absolute judges. Instead of assigning direct difficulty values, LLMs compare pairs of items and determine which item is more difficult for a target population. [Raina and Gales \(2024\)](#), for example, found that comparative prompting produced substantially stronger difficulty rankings for multiple-choice reading comprehension items than direct absolute scoring. However, their work focused on relative ranking counts rather than psychometric calibration.

Despite growing interest in LLM-based assessment methods, limited research has examined whether LLM-generated comparative judgments can be incorporated into an anchored psychometric calibration framework. The present study addresses

this gap using TIMSS Grade 4 mathematics items and a fixed-anchor Bradley-Terry calibration approach.

3 Methods

The study used Grade 4 mathematics items from TIMSS, including 80 previously calibrated trend items that served as fixed anchors (difficulty range: -1.247 to 1.592) and 101 newly developed TIMSS 2023 items targeted for difficulty estimation (range: -2.979 to 1.797). Operational 2023 item difficulties were withheld during estimation and used only for validation.

The study followed a two-stage design. A pilot calibration used 12 anchors selected to provide approximately uniform coverage across the trend-item difficulty range. This pilot was used to evaluate whether a sparse anchor design could support LLM-based calibration and to identify items requiring additional information. Results from the pilot indicated that some items exhibited complete win-loss patterns, extreme Bradley-Terry estimates, or weak graph connectivity. These findings motivated an enhanced calibration design.

Item difficulty information was obtained through comparative judgment. For each comparison, the LLM was presented with two mathematics items and asked to determine which item would be more difficult for a typical Grade 4 student. Three judge prompts were used: Mathematics Content Expert Judge, Student Probability Judge, and Cognitive Load Judge. The full prompts for each judge are provided in the Appendix. Each judge evaluated the same comparison pairs independently, and responses were recorded as binary outcomes indicating which item was judged more difficult.

In the enhanced design, each new 2023 item was compared against 24 fixed trend items. Anchor items were selected using stratified sampling across the available difficulty range to ensure that each target item was compared against relatively easy, moderate, and difficult reference items. Each new item was also compared with 10 randomly selected new items to improve graph connectivity and provide information about the relative ordering among new items. In addition, 31 items identified as unstable in the pilot calibration received eight additional anchor comparisons and eight additional bridge comparisons. Table A1 summarizes the pilot and enhanced comparison designs.

LLM judgments were generated using Qwen3-

8B with deterministic decoding. The model was prompted using chat-formatted messages with thinking disabled. Generation used a maximum of 1,024 new tokens, temperature set to 0, and no sampling. A fixed random seed was used for comparison sampling and reproducibility.

Pairwise judgments were modeled using the Bradley-Terry model. Let $Y_{ij} = 1$ indicate that item i was judged more difficult than item j . The probability of this outcome is modeled as:

$$P(Y_{ij} = 1) = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}, \quad (1)$$

where

$$\eta_{ij} = \theta_i - \theta_j, \quad (2)$$

which represents the difference between the latent difficulty parameters associated with the two items.

Unlike a conventional Bradley-Terry model in which all item parameters are estimated, the present study treated trend items as fixed anchors. Known TIMSS trend-item difficulty estimates were incorporated as fixed offsets to identify the Bradley-Terry scale and provide reference locations for estimating the relative difficulty of newly developed items. Accordingly, anchor-item parameters were not estimated during model fitting.

For a comparison involving a new item i and an anchor item a , the linear predictor can be expressed as

$$\eta_{ia} = \theta_i - b_a, \quad (3)$$

where θ_i is the unknown difficulty parameter of the new item and b_a is the fixed difficulty parameter of the anchor item. More generally, for any comparison, the linear predictor was defined as the difference between the estimated parameters of non-anchor items and the fixed difficulty values of anchor items when present. Bradley-Terry estimation was performed using the BradleyTerry2 package in R.

Estimated item difficulties were evaluated against operational TIMSS 2023 difficulty parameters that were withheld during model estimation. Agreement was assessed using Pearson and Spearman correlations, difficulty-band classification based on tertiles, adjacent-band agreement, and pairwise ordering accuracy.

4 Results

The three judges produced 11,607 pairwise judgments in total, spanning around 3,800 unique item

pairs per judge. Of these, 7,272 were comparisons of a new item against a fixed trend anchor, 2,856 were bridge comparisons between new items, and the remaining 1,479 were additional anchor and bridge comparisons assigned to the 31 items flagged as unstable in the pilot. No anchor-anchor comparisons were used, consistent with the fixed-offset specification. Unlike the pilot, the enhanced design produced no items with complete win or loss patterns: new-item win rates ranged from 0.016 to 0.943, so all 101 items received finite, well-identified Bradley-Terry estimates.

In addition to comparative difficulty judgments, the LLM also provided a self-reported confidence value and a perceived difference magnitude classification (“slight,” “moderate,” or “large”) for each comparison. Confidence values showed minimal variation, with nearly all responses concentrated at 0.85, limiting their usefulness for calibration analysis. In contrast, perceived difference labels showed meaningful variation in ordering accuracy. Comparisons labeled as “moderate” differences agreed with the operational item ordering 69.5% of the time, whereas comparisons labeled as “slight” differences were correct only 54.0% of the time. However, the perceived difference labels themselves were essentially uncorrelated with the operational difficulty gap ($r = -.02$), suggesting that these labels reflected the LLM’s perceived decisional certainty rather than calibrated estimates of latent difficulty separation.

Across all 101 TIMSS 2023 items, the pooled estimates correlated with the withheld operational b -parameters at $r = .484$ (Spearman $\rho = .484$). The three-judge ensemble was essentially identical ($r = .485$, $\rho = .486$). Agreement was also broadly similar across the three individual judges, with the Mathematics Content Expert Judge producing the strongest correspondence with operational difficulty. Table 1 summarizes correlations across judges and aggregation methods. The ensemble did not substantially exceed the strongest individual judge, suggesting that combining judges contributed robustness rather than additional predictive accuracy. At face value, these results represent moderate recovery of the operational difficulty ordering, but the aggregate correlations are inflated by the structure of the validation set and are best interpreted alongside the subgroup results below.

For items whose operational difficulties fell within the anchor range ($n = 69$), the relationship between estimated and operational difficulty

Model	Pearson (r)	Spearman (ρ)
Pooled	.484	.484
Ensemble	.485	.486
Math Expert	.489	.509
Student Probability	.482	.498
Cognitive Load	.456	.437

Table 1: Agreement between Bradley-Terry difficulty estimates and operational parameters across judges and aggregation methods.

was moderate ($r = .372$, $\rho = .361$). In contrast, recovery was weak for items below the anchor range ($n = 31$, $r = -.041$, $\rho \approx .00$); and only one item fell above the anchor range, preventing estimation of a meaningful correlation. Figure 1 illustrates this pattern. Items below the anchor range clustered at the lower end of both the operational and estimated scales, whereas in-range items occupied the central portion of the scale. Consequently, the overall correlation across all items was partially driven by between-group separation rather than precise recovery of within-range variation. These findings suggest that the anchored comparative judgment framework was more effective for interpolating within the range represented by the fixed anchors than for extrapolating beyond that range.

Difficulty-band classification analyses produced similar results. When items were divided into tertiles based on operational difficulty, exact-band classification accuracy was 54.5%, substantially above the 33% chance baseline. In addition, 90.1% of items were classified into either the correct difficulty band or an adjacent band. Classification accuracy was strongest for the lowest and highest tertiles (55.9% and 60.6%, respectively) and weakest for the middle tertile (47.1%), consistent with the greater difficulty of distinguishing items near the center of the scale. Pairwise ordering accuracy, defined as the proportion of item pairs whose relative ordering was correctly recovered, was 66.8%, indicating that the comparative judgment framework preserved much of the operational difficulty ordering even when exact parameter recovery was imperfect.

5 Discussion

The present study examined whether LLM-generated comparative judgments could support early-stage estimation of item difficulty within an anchored psychometric framework. Unlike several recent approaches that incorporate empirical

response information into prompts, such as student performance statistics or explicit psychometric guidance (e.g., Jain et al., 2025), the current framework relied exclusively on item content and pairwise comparative judgments. From an operational perspective, this is particularly relevant during item development and pretesting, when empirical response data are not yet available.

Results showed that pairwise judgments produced by the LLM contained meaningful information about operational TIMSS item difficulty, with moderate correspondence observed between Bradley-Terry estimates and operational item parameters. Performance depended strongly on anchor coverage: agreement was moderate for items within the anchor range and weak for items outside it, indicating that the framework is more effective for interpolation than extrapolation. These findings suggest that comparative judgment may provide a useful supplementary source of difficulty information prior to field testing.

The enhanced comparison design substantially improved estimation stability relative to the pilot calibration. Expanding the anchor pool, introducing bridge comparisons, and adding targeted comparisons for unstable items eliminated complete win-loss patterns and produced finite estimates for all items. This highlights the importance of graph connectivity and anchor coverage in comparative calibration frameworks and suggests that the structure of the comparison network may be as important as the choice of LLM judge itself.

Limitations and Future Research

Several limitations should be acknowledged. First, the study relied on a single, relatively small open model (Qwen3-8B), and results may differ for larger or differently aligned models. Second, the study focused exclusively on TIMSS Grade 4 mathematics items, limiting generalizability to other subjects and grade levels. Third, operational item difficulty is itself population-dependent, meaning that part of the residual disagreement likely reflects differences between item features and student-item interaction rather than judgment error alone (Yaneva et al., 2024). Finally, the Bradley-Terry estimates represent locations on a judgment-derived latent scale rather than direct operational IRT difficulty parameters. Accordingly, the framework should be viewed as a supplementary pretesting tool rather than a replacement for operational psy-

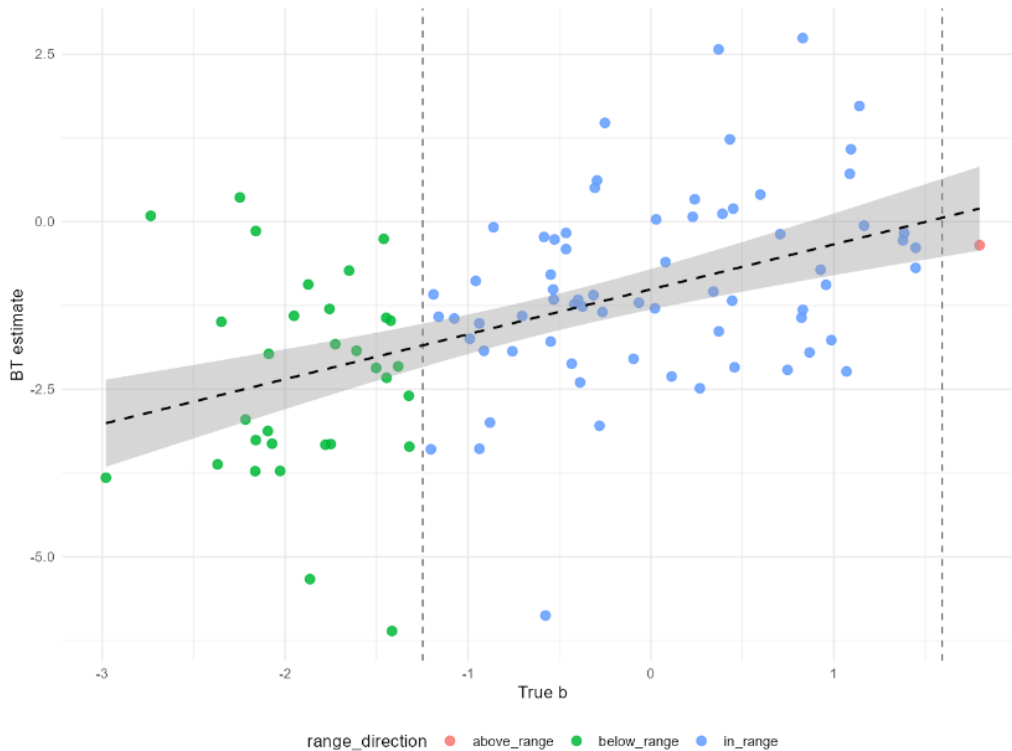


Figure 1: Relationship between operational TIMSS difficulty parameters and pooled Bradley-Terry estimates.

chometric calibration.

Despite these limitations, the study demonstrates that LLM-generated comparative judgments can be integrated with fixed-anchor psychometric scaling procedures to provide meaningful preliminary information about item difficulty before field-test data are available. Several directions would clarify and extend this finding. Comparing additional models, especially utilizing larger proprietary models, would establish whether the moderate ceiling observed here is intrinsic to the approach or an artifact of a small model. Widening and densifying the anchor set, particularly at the extremes, offers the most direct route to reducing extrapolation failures and improving in-range recovery, while developing data-driven screening for out-of-range items would make the method deployable without the prior knowledge of true difficulty. Extending the framework to other subjects and grades, and benchmarking LLM judges against human comparative judges within the same anchored design, would further describe where AI-supported comparative judgment can responsibly contribute.

References

- S. AlKhuzaei, F. Grasso, T. R. Payne, and V. Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Y. Attali, L. Saldivia, C. Jackson, F. Schuppan, and W. Wanamaker. 2014. Estimating item difficulty with comparative judgments. *ETS Research Report Series*, 2014(2):1–8.
- Isaac I Bejar. 1983. Subject matter experts’ assessment of item statistics. *Applied Psychological Measurement*, 7(3):303–310.
- U. Bezirhan and M. von Davier. 2025. Ai-based classification of timss items for framework alignment. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 134–141.
- R. A. Bradley and M. E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- R. K. Hambleton, S. G. Sireci, H. Swaminathan, D. Xing, and S. Rizavi. 2003. Anchor-based methods for judgmentally estimating item difficulty parameters. Technical report, LSAC Research Report Series.
- Y. Jain, J. Hollander, A. He, S. Tang, L. Zhang, and J. Sabatini. 2025. Exploring the potential of large

- language models for estimating the reading comprehension question difficulty. In *International Conference on Human-Computer Interaction*, pages 202–213. Springer Nature Switzerland.
- R. Kapoor, S. T. Truong, N. Haber, M. A. Ruiz-Primo, and B. W. Domingue. 2025. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.
- M. Mansoor, S. Imran, A. Tayyab, and R. Sarfraz. 2024. Expert prediction versus difficulty index measured by psychometric analysis; a mixed method study interpreted through diagnostic judgment by cognitive modeling framework. *Journal of University College of Medicine and Dentistry*, pages 74–80.
- J. W. Park, S. J. Park, H. S. Won, and K. M. 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.
- V. Raina and M. Gales. 2024. Question difficulty ranking for multiple-choice reading comprehension. *arXiv preprint arXiv:2404.10704*.
- A. C. Rogoz and R. T. Ionescu. 2024. Unibucllm: Harnessing llms for automated prediction of item 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 493–502.
- L. L. Thurstone. 1927. [A law of comparative judgment](#). *Psychological Review*, 34(4):273–286.
- H. Turner and D. Firth. 2012. Bradley-terry models in r: The bradleyterry2 package. *Journal of Statistical Software*, 48:1–21.
- V. Yaneva, K. North, P. Baldwin, L.A. Ha, S. Rezayi, Y. Zhou, S.R. Choudhury, P. Harik, and B. 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.

A Appendix

Component	Pilot	Enhanced
Anchors	12	80
New items	101	101
Anchor comparisons	12	24
Bridge comparisons	0	10
Adaptive comparisons	No	Yes

Table A1: Comparison designs for the pilot and enhanced calibration

LLM judge response format and prompts

All judges were instructed to respond only using the following JSON format:

```
{
  "harder_item": "A_or_B",
  "difference": "slight|moderate|large",
  "confidence": 0.0-1.0,
  "reason": "brief"
}
```

Mathematics Content Expert Judge You are evaluating comparative difficulty for TIMSS Grade 4 mathematics assessment items. Your task is to estimate which item would have the lower probability of correct response among actual Grade 4 students. Focus primarily on mathematical difficulty rather than superficial appearance. Consider conceptual prerequisites, procedural complexity, number of reasoning steps, abstraction level, representation complexity, working-memory demands, common student misconceptions, and coordination of multiple mathematical ideas. Ignore wording and style unless they change mathematical demand. Many item pairs differ only slightly in difficulty. Avoid exaggerating small differences.

Student Probability Judge You are estimating comparative item difficulty for TIMSS Grade 4 mathematics items. Judge which item would have the lower probability of correct response among a large representative sample of Grade 4 students. Base judgments on expected empirical student performance. Consider conceptual prerequisites, procedural demands, number of required operations, representation complexity, likely misconceptions, reading and interpretation burden, and reasoning coordination demands. Many item pairs differ only slightly. Use conservative, calibrated comparisons.

Cognitive Load Judge You are evaluating which TIMSS Grade 4 mathematics item would impose greater cognitive demand on actual Grade 4 students. Estimate which item would produce lower student success probability. Focus on working-memory load, multi-step reasoning, information integration, intermediate results, abstraction, representation switching, and hidden conceptual traps. Many item pairs are close in difficulty. Do not overstate small differences. Ignore stylistic wording differences unless they affect cognitive demand.

Table A2: LLM judge response format and prompts