

Using LLM Judges' Paired Comparisons to Estimate Mathematics Item Difficulty

Lanrong Li, Mohammed A. A. Abulela, Matthew Gushta

MetaMetrics, Inc.

{jli, mabulela, mgushta}@metametrics.com

Abstract

Estimating item difficulty without collecting field testing data has been a long sought-after goal in educational measurement. We prompted large language models (LLMs) to compare mathematics items in pairs to estimate their difficulty. Results showed strong correlations between difficulty based on one LLM judge's paired comparisons and empirical item difficulty.

1 Introduction

Estimating item difficulty without collecting field test data has been a long sought-after goal in educational measurement. This line of research is not only motivated by the potential of substantially reducing the cost and time for estimating the difficulty of newly developed items, but also by the needs of personalized learning and adaptive assessments. A variety of approaches have been proposed to estimate item difficulty without using participants' item response data, spanning using content experts' judgments (e.g., [Bejar, 1981](#)), applying traditional statistical models (e.g., [Rupp, 2011](#)), machine learning (e.g., [Cai et al., 2026](#)), or deep learning models (e.g., [Huang et al., 2017](#)) to item features, leveraging large language models (LLMs; e.g., [He et al., 2024](#); [Maeda, 2025](#)), or combining multiple methods ([Razavi & Powers, 2025](#)). However, each approach has its limitations. For example, using content experts' judgments is costly and time consuming, using feature-based approaches requires pre-extracted item features and known item difficulties, while using LLMs as respondents (e.g., [He et al., 2024](#)) or direct difficulty estimators (e.g., [Razavi & Powers, 2025](#)) has achieved limited success.

Recently, LLMs have been increasingly used as evaluators to make judgments (e.g., [Liu et al., 2024](#); [Liusie et al., 2024](#); [Qin et al., 2023](#)) on natural language processing tasks. Paired comparison ([Thurstone, 1927](#)), in which two items are presented and compared at a time, has been shown to be more effective than direct estimation with LLMs (e.g., [Liusie et al., 2024](#)). Paired comparison data can be analyzed using the Bradley-Terry model ([Bradley & Terry, 1952](#)) to obtain quantitative measures of the attribute of interest for all evaluated candidates. In educational measurement, however, few studies used this approach to estimate item difficulty. One exception is the study by [Li and Gushta \(2026a\)](#), which investigated the performance of LLMs in making paired comparisons to estimate English reading item difficulty. Two other relevant studies are [Ballon et al. \(2025\)](#) and [Kolesnikova et al. \(2026\)](#), but neither controlled for model position bias ([Liusie et al., 2024](#)) nor used items containing both text and images. This study aims to use the paired comparisons generated by multimodal LLMs to estimate the difficulty of mathematics items that contain text, mathematical symbols, or images.

2 Background

Earlier methods for estimating item difficulty without using participants' item response data include asking subject matter experts to estimate item difficulty directly (e.g., [Lorge & Kruglov, 1952](#); [Bejar, 1983](#)) or make comparative judgments ([Attali et al., 2014](#)), and applying regression models to extracted item features ([Rupp et al., 2001](#)). With the advancement in data science, the use of machine learning, deep learning, and natural language processing in item difficulty prediction has been increasing (see the review by [Alkhuzaey et al., 2024](#)). The emergence of LLMs attracted a

surge of interest in their potential to predict item difficulty. Some researchers prompted them to predict item difficulty directly (e.g., [Kolesnikova et al., 2026](#)), whereas others used them as respondents to generate item responses for subsequent item calibration (e.g., [Maeda, 2025](#)). In these studies, empirical item difficulties based on human participants' item response data were often used as the ground truth. The correlations between predicted and empirical item difficulty observed across approaches ranged from weak to strong. For example, [Razavi and Powers \(2025\)](#) reported correlations ranging from .13 to .55 across individual grade levels and a correlation of .83 when using all 2,564 mathematics items developed for Grades K–5 students.

LLMs, due to their remarkable capability of understanding human languages and performing tasks that require reasoning, a new paradigm of LLM-as-a-judge has emerged ([Zheng et al., 2023](#)), in which LLMs are used as judges to assess the quality of the outputs of other AI models or systems. Two methods are often used when employing LLMs as judges: one is to instruct the LLM to provide a rating directly based on predefined criteria, and the other is to instruct it to make pairwise comparisons. Pairwise comparison is a method first proposed by [Thurstone \(1927\)](#) in psychology and was later applied in many other fields. It is simple, intuitive, easy to implement. Paired comparison ratings can be analyzed by the Bradley-Terry model ([Bradley & Terry, 1952](#)) to generate estimates of target attributes on an interval scale. Although as the number of candidates to be compared increases, the number of all possible comparisons increases quadratically, not all possible pairs need to be used. [Jones and Davies \(2023\)](#) suggest that 10 comparisons per candidate is sufficient for most cases. [Li and Gushta \(2026a\)](#) has shown that as few as 10 comparisons per item can be used to obtain stable and useful item difficulty estimates. The paired comparison method has also been shown to be more effective than the direct rating method (e.g., [Liu et al., 2024](#); [Liusie et al., 2024](#); [Qin et al., 2023](#)).

While using LLMs as judges, their various biases have also been revealed (e.g., [Wang, 2023](#)). For example, it has been found that LLMs show position bias during pairwise comparisons (e.g., [Liusie et al., 2024](#)), in which they tend to select the

first or the second item in a pair more frequently regardless of the amount of the attributes of the two items. The swapping method ([Zheng et al., 2023](#)), in which the presentation order of the two items is reversed, has proven to be a simple and effective technique for mitigating the position bias. Prior research (e.g., [Loya et al., 2023](#)) also found that LLMs are sensitive to prompt variations. [Wei et al. \(2022\)](#) demonstrated that chain-of-thought prompting resulted significantly enhanced LLMs performance on complex reasoning tasks. However, whether prompt variations similarly influence the performance of LLMs in paired comparisons remains an open empirical question.

In educational measurement, several studies have used the paired comparison method with LLMs as judges to estimate item difficulty. For example, [Ballon et al. \(2025\)](#) used OpenAI's o3 and Gemini 2.5 Pro to make paired comparisons on items that measure advanced math (34 items), English reading (787 items), and algebra (1876 items), respectively. They found the difficulty estimates based on the paired comparisons of Gemini-2.5-pro correlated with empirical difficulty at .44 for the advanced math items and .53 for English reading items, while OpenAI o3 generated weaker correlations. Although the correlations were not strong, this study demonstrated the promise of the paired comparison approach for estimating item difficulty. However, the study did not account for LLM position bias or provide sufficient details on how the empirical item difficulties were obtained.

[Kolesnikova et al. \(2026\)](#) examined the performance of three LLMs, GPT-4o, DeepSeek-Chat, and Qwen3-235b, in direct difficulty estimation and pairwise comparisons of 360 mathematics items measuring six domains¹ of elementary school arithmetic operations and applications. Results showed an average Spearman correlation of .73 across domains and LLMs, using paired comparisons with a zero-shot prompting and a hard decision rule (i.e., based on explicit outputs of the LLMs). They also found that pairwise judgments yielded better correlations than absolute judgments by LLMs, and that incorporating few-shot prompting or token-level probability failed to outperform zero-shot prompting or the hard decision rule. However, as with the study by [Ballon et al. \(2025\)](#), the study did not control for position

¹ Paired comparisons were made separately for each domain.

bias of the LLMs, so it was possible that the reported correlations were inflated.

Li and Gushta (2026a) selected five LLMs and prompted them to serve as judges and make pairwise comparisons of the difficulty of 40 Grade 5 English reading items. They examined within-model and between-model reliability of the ratings and the effects of model temperature and number of comparisons per item on rating accuracy and item difficulty recovery. Results showed that low temperature led to high between- and within-model reliability. Item difficulty estimates based on LLM judges' paired comparisons correlated moderately to strongly with empirical item difficulty, with the highest coefficient at .62 based on the ratings by Gemini-3.1-flash. This study provided many useful implications for practitioners, but it used a small set of reading items and a single prompt. It remains to be seen how LLMs would perform with items from a different content domain, a larger item set, and different prompts.

This study aims to examine the performance of LLM in making paired comparisons to estimate the difficulty of mathematics items developed for Grades 3-8 students. We selected three LLMs and designed several prompts to identify the best performing model and the most effective prompt. We also investigated factors that affect the rating accuracy of LLMs by using empirical item difficulty as the ground truth. Specifically, we intend to answer the following research questions:

1. How well are mathematics item difficulties estimated using paired comparisons of LLMs?
2. Does prompt design affect LLMs' performance in making paired comparisons?
3. Are there any item-pair related features that affect the rating accuracy of the LLMs in paired comparisons?

3 Methods

3.1 Mathematics Items

A total of 178 mathematics items developed for students in Grades 3-8 in the U.S. were used for the study. These items were from an empirical study in which 11,786 students in Grades 3-8 took a fixed-form mathematics test. Test lengths ranged from 30 to 36 items, depending on grade levels. Common items were designed for vertical linking, resulting in 178 unique items in the pool. These items were intended to measure the following content

domains: (1) Number Sense; (2) Numerical Operations; (3) Algebra and Algebraic Thinking; (4) Geometry; (5) Measurement; and (6) Data Analysis, Statistics, and Probability (MetaMetrics, 2022). All items are multiple-choice, each consisting of a stem and four options. The empirical difficulty estimates of the items were obtained by applying the Rasch model (Rasch, 1960/1980) to students' item response data pooled across grades. Consequently, the estimated difficulties of all items are on the same scale. A sample item is shown below:

Felipe purchased a suit on sale for \$131.25. The sale price was 30% off of the original price. Which was the original price of the suit?

- A. \$91.88
- B. \$170.63
- C. \$187.50 (correct option)
- D. \$437.50

The sample item above only contains text and numbers. However, most items have mathematical symbols, and about 39% of the 178 items have images in the stem, the options, or both.

3.2 Large Language Models

Although a wide range of LLMs are available, due to the characteristics of the items used in this study, we applied the following criteria when selecting LLMs: (1) the model is capable of processing both text and image data; (2) the model has strong mathematical reasoning capability; and (3) the model is fast and cost-effective. Using these criteria, we selected three LLMs: Gemini-3.5-flash, Claude-sonnet-4.6, and GPT-5.4.

3.3 Experimental Design

For Gemini-3.5-flash and Claude-sonnet-4.6, the temperature was set to 0 to minimize randomness in their responses and maximize within-model rating reliability. GPT-5.4 does not allow temperature setting so the default 1 was used. To control the position bias of LLMs, we used the swapping method (Zheng et al., 2023) to switch the positions of the items in each pair and instructed the LLM to compare the same pair in both orders. If the selected items from the two comparisons for the same item pair were the same, it was recorded as the final rating for that pair; otherwise, it was a tie. The tie also reflects models' position bias.

With N items, the total number of unique item pairs is $N*(N-1)/2$. Based on the findings by Li and

Gushta (2026a) and following the guidelines by Jones and Davies (2023), 10 comparisons per item were executed. Only 33 Grade 5 items were used to study the prompt sensitivity of LLMs and identify the best-performing model and the most effective prompt. Grade 5 items were selected because they are considered most representative of the entire Grades 3-8 item pool.

All models had the same system prompt or persona (see the appendix). To examine prompt sensitivity of LLMs, we designed three user prompts: (1) a prompt that simply instructed the LLM to compare the difficulty of two items in each pair and choose the one that is more difficult for students; (2) a prompt that includes everything in Prompt 1 but added vague evaluation criteria (i.e., criterion with names but no descriptions) for the model to consider when comparing item difficulty; and (3) a prompt containing detailed evaluation criteria, each accompanied by a description and illustrative examples (see the appendix). For each item, the stem, options, and key were provided as input to the LLM. For items with images, the images files (save in PNG format) were included as part of the input. Each LLM served as a judge and was instructed to complete all paired comparisons for each prompt. All models were configured to operate at a “high” reasoning or thinking level.

Paired comparison data using Grade 5 items were collected between June 11 and June 14, 2026. The selected model and prompt were then used on all 178 items. The paired comparison data for all items was collected between August 28 and August 30, 2026.

3.4 Data Analysis

After data collection, the paired comparison data was analyzed by applying the Bradley-Terry model (Bradley & Terry, 1952) using Facets (Linacre, 2023a). Using the model, the probability of an item i being judged as more difficult than item j is:

$$Prob(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}, \quad (1)$$

where β_i and β_j are item difficulty parameters on the logit scale. Joint maximum likelihood method was implemented in Facets to estimate item parameters, assuming local independence among pairwise comparisons.

After obtaining initial item difficulty estimates from Facets, we observed that they had a large spread, which is often caused by data sparseness, a Guttman-like ordering of the items, or the presence

of extremely easy or difficult items (Linacre, 2023b). For a fair comparison between prompts and LLMs, we applied linear transformation to item difficulty estimates based LLM judges’ paired comparisons such that the transformed estimates had the same mean and standard deviation as the empirical difficulty estimates.

To evaluate item difficulty recovery, we computed the Pearson correlation between LLM-paired-comparison-based difficulty estimates and empirical difficulty estimates of 33 Grade 5 items and the root mean squared error (RMSE) of LLM-derived difficulty. The correlations and RMSEs were compared across prompts and models to determine the most effective prompt and best-performing model. The prompt or the model yielding the strongest correlation and the smallest RMSE was designated as optimal. The best-performing LLM with the most effective prompt was then used with the entire 178 items in the Grades 3-8 item pool.

To evaluate prompt sensitivity of the LLMs, we compared item difficulty recovery across prompts and then performed a chi-square independence test on the rating accuracy of each model with different prompts. Rating accuracy was computed by dividing the number of times the correct item was selected in a pair (judged by their empirical difficulty) by the total number of item pairs compared. When a tie occurred, the accuracy rate for the item was halved. Additionally, we also computed the tie rate to record position bias of the models. The analyses were performed for Grade 5 items only.

To answer the third research question, we fit a logistic regression model to the paired comparisons of all 178 items using rating accuracy as the outcome variable and three identified predictors. The rating accuracy had two values, with 0 indicating the wrong item was selected and 1 otherwise. The three predictors are: (1) whether the two items were developed for the same grade level; (2) whether at least one item in the pair has images in its stem or options, and (3) whether the two items measure the same content strand. We hypothesized that the first predictor would contribute most to rating accuracy, given that difficulty of items is typically strongly correlated with their associated grade level. The effect of image on the rating accuracy was examined because preliminary testing revealed that certain LLMs produced incorrect answers to items with images in the stem

or options during exploration. Difference in content domains could be another factor affecting accuracy because a previous study (Li & Gushta, 2026b) found that construct heterogeneity can influence human judges' paired comparisons of reading items.

4 Results

Table 1 shows item difficulty recovery results using 33 Grade 5 items. Among the three LLMs, Gemini-3.5-flash yielded the highest correlations between LLM-derived item difficulty and their empirical counterparts, with coefficients around .70 across prompts. GPT-5.4 and Claude-4.6-sonnet generated lower correlations between the two sets of item difficulties, with coefficients around .60. The 95% confidence intervals of the correlation coefficients are wide, reflecting the small number of items in the analyses. With regard to RMSE, Gemini-3.5-flash was also the winner because of its lowest RMSEs in the item difficulty estimate among the three LLMs. The other two LLMs had comparable RMSEs. Figure 1 shows the scatter plot between LLM-derived difficulty and empirical difficulty for 178 Grades 3-8 items based on the paired comparisons of Gemini-3.5-flash with a simple prompt. A strong, positive, and linear relationship was observed between the two sets of item difficulties. The correlation coefficient was .80 with 95% confidence interval between .74 and .85. The RMSE of the item difficulty estimates was .77.

In terms of prompt sensitivity of the LLMs, there appears to be a weak interaction effect between

LLM	Prompt	r [95% CI]	RMSE
Claude-4.6-sonnet	Simple	.62 [.35, .79]	.87
	Vague	.57 [.28, .76]	.93
	Detailed	.56 [.27, .76]	.93
Gemini-3.5-flash	Simple	.70 [.48, .84]	.77
	Vague	.69 [.46, .84]	.78
	Detailed	.71 [.48, .85]	.76
GPT-5.4	Simple	.59 [.31, .77]	.91
	Vague	.62 [.36, .80]	.86
	Detailed	.56 [.26, .76]	.94

Table 1: Item difficulty recovery of Grade 5 items.

prompt type and LLM: Gemini-3.5-flash appeared insensitive to different prompts, but a simple prompt led to marginally better performance of Claude-4.6-sonnet, while a prompt with vague

evaluation criteria for item difficulty comparison

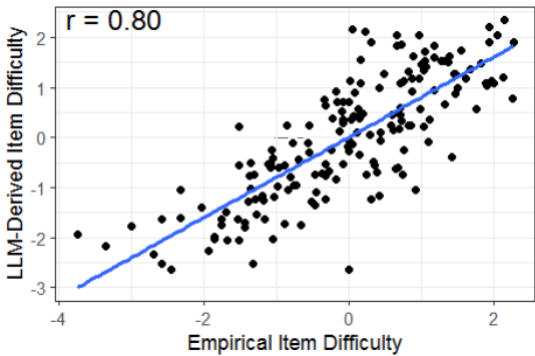


Figure 1: Scatter plot of LLM-derived and empirical difficulty of 178 Grades 3-8 mathematics items.

resulted in slightly better performance by GPT-5.4.

Table 2 shows the position bias (indicated by tie rate) and rating accuracy of the LLMs and the chi-square statistic for rating accuracy. Gemini-3.5-flash showed the least amount of position bias among the three LLMs, followed closely by GPT-5.4, while Claude-4.6-Sonnet had the highest position bias. Similarly, Gemini-3.5-flash also had the highest accuracy rate in its ratings, followed by GPT-5.4 and Claude-4.6-Sonnet. The chi-square independence tests were not significant for any prompts, suggesting that prompt type had little effect on the rating accuracy of the LLMs.

LLM	Prompt	Tie %	Accuracy %	χ^2
Claude-4.6-sonnet	Simple	24	56	.76
	Vague	22	55	
	Detailed	18	58	
Gemini-3.5-flash	Simple	9	66	.04
	Vague	11	66	
	Detailed	11	67	
GPT-5.4	Simple	11	62	.14
	Vague	13	63	
	Detailed	10	63	

Table 2: Position bias (tie rate) and rating accuracy of LLMs across prompts based on 33 Grade 5 items.

The effects of the three selected factors on the rating accuracy of Gemini3.5-flash with a simple prompt and 178 mathematics items are displayed in Table 3. Two factors, whether two items have the same assigned grade levels and whether at least one item has images in a pair had statistically significant effects on the rating accuracy. *grade_same* had a significant negative effect, indicating that same grade levels resulted lower

rating accuracy than different grade levels. When at least one item has images in its stem, options, or both, it had a positive effect on the rating accuracy. Whether two items in a pair measure the same content strand had little effect on the rating accuracy, as indicated by the nonsignificant coefficient ($p = .664$) and the odds ratio close to 1.

Variable	$\hat{\beta}$	SE	$\exp(\hat{\beta})$	Z
Intercept	1.08	0.11	2.93	9.48***
grade_same	-0.55	0.15	0.58	-3.65***
with_image	0.41	0.13	1.50	3.24**
strand_same	0.07	0.15	1.07	0.44

Note. *** $p < .001$; ** $p < .01$.

Table 3: Results of logistic regression analysis of rating accuracy of Gemini-3.5-flash.

5 Discussion

This study investigates the effectiveness of using LLM judges' paired comparisons to estimate mathematics item difficulty. The results for the 33 Grade 5 mathematics items showed that the correlations between LLM-paired-comparison-based difficulty estimates and empirical difficulty estimates were moderate to strong, with Gemini-flahs-3.5 achieving the best performance in item difficulty recovery among the three LLMs. [Ballon et al. \(2025\)](#) also found that Gemini-2.5-Pro performed better than OpenAI's o3. The better performance of Gemini might be attributed to its stronger capability of processing image data ([Kokalj, 2026](#)). When using the paired comparisons of Gemini-flash-3.5 on the entire 178 Grades 3-8 mathematics items with a simple prompt, the correlation coefficient reached .80. The correlation was quite strong, only slightly weaker than the highest correlation reported in [Razavi and Powers' \(2025\)](#) study, but that study used 2,564 mathematics items and considerable effort and resources were invested in extracting item features. Also, it is also worth noting that only 10 comparisons per item were used for the 178 items.

Regarding the prompt sensitivity of the LLMs, the patterns were slightly different among the models using item difficulty correlations and RMSE as the criteria. For Gemini-3.5-flash, minimal difference in correlation or RMSE were observed among prompts. For Claude-4.6-sonnet, the simple prompt resulted in better item difficulty recovery, while for GPT-5.4, a prompt with vague evaluation criteria demonstrated a marginal

advantage. We did not find statistically significant differences among the three prompts in the rating accuracy across LLMs. Overall, no strong evidence was found to support that providing detailed comparison criteria in the prompt increases the performance of LLMs in paired comparison of mathematics items. The results were probably due to the fact that the reasoning levels of all the large language models used were already high. That is, even without providing any evaluation criteria, the LLMs already used a set of criteria similar to the criteria provided in the other two prompts to compare item difficulties.

In terms of item features that contribute to the rating accuracy of the LLMs, two significant predictors were identified: whether the intended item grade levels are the same and whether at least one item in a pair contains images. Given the strong correlations observed between LLM-based and empirical difficulty, it was expected to see whether item grade levels are the same in a pair was a significant predictor. Items with the same grade level tend to be more similar in difficulty than items with different grade levels. The higher similarity in difficulty for items with the same grade level should have made it harder for Gemini-3.5-flash to determine the correct item in a pair than items with larger differences in difficulty. The second factor, whether at least one item has an image in its stem or options, had a positive effect on the rating accuracy, suggesting that when this condition is true, Gemini-3.5-flash has better performance in paired comparison of mathematics item difficulty. Whether the content strands are the same for items in a pair had little effect on the rating accuracy. This was likely because the measured strands have a clear progression in curriculum and instruction; mathematics skills and knowledge taught later to students inherently build on foundational concepts introduced earlier. Therefore, whether two items in a pair measure the same content strands had little impact on the rating accuracy of the model.

The findings of the study has important implications for practitioners. It shows that using paired comparison with LLMs can generate difficulty estimates of mathematics items that strongly correlated with empirical item difficulties. These correlations are strong enough to be useful for low-stakes classroom assessment, personalized learning, and adaptive testing. The correlations are higher than many of the correlations found using alternative human-response-data-free approaches

to item difficulty estimation and comparable to those using resource-intensive data-driven approaches. Additionally, compared to alternative approaches, it is more efficient, cost-effective, and easier to implement. Furthermore, using LLMs as judges with the paired comparison method does not require items with existing difficulty. This is especially valuable for small or new testing programs where it is not feasible to conduct field testing to obtain initial estimates of item difficulty.

This study has several limitations. First, in order to use LLMs that were roughly comparable in performance, we chose the three models. However, due to the unique features of the models, their temperatures settings differed, making it difficult to ensure a fair comparison across models. Second, our model selection was constrained by cost efficiency. Future research with access to greater computation resources should evaluate larger or more advanced models, which may yield more accurate item difficulty estimates. Third, we only implemented 10 comparisons per item. Li and Gushta's (2026a) study found little difference in item difficulty recovery between 10 and 15 comparisons per item. However, that study was based on a small set of items. It is possible that more comparisons per item would result in better item difficulty estimation for a larger item set. This could be an experiment design factor for future studies. Fourth, we did not examine between- or within-model reliability of the ratings because we assumed that the findings from Li and Gushta (2026a) study based on reading items generalize to mathematics items. Given the critical role of reliability in paired comparison, evaluating this assumption is particularly important when the model temperature is high. Fifth, we did not prompt LLMs to generate justifications or rationale for their choices. Eliciting qualitative explanations could yield valuable insights into LLM reasoning during pairwise evaluations, ultimately informing strategies to optimize prediction accuracy.

References

- AlKhuyaey, 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. <https://doi.org/10.1007/s40593-023-00362-1>
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). *Estimating item difficulty with comparative judgments* (ETS Research Report No. RR-14-39). Princeton, NJ: Educational Testing Service. <https://files.eric.ed.gov/fulltext/EJ1109277.pdf>
- Ballon, M., Algaba, A., Verbeken, B., & Ginis, V. (2025). *Estimating problem difficulty without ground truth using large language model comparisons*. arXiv. <https://doi.org/10.48550/arXiv.2512.14220>
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7(3), 303-310. <https://doi.org/10.1002/j.2333-8504.1981.tb01274.x>
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324. <https://doi.org/10.2307/2334029>
- Cai, H., Yan, X., Chuang, P. L., Pan, Y., & Huo, M. (2026). What makes listening comprehension difficult?: A feature-based machine learning approach to understanding item difficulty. *Applied Linguistics*, 47(2), 406-428. <https://doi.org/10.1093/applin/amaf079>
- He-Yueya, Joy, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W. Domingue, Emma Brunskill, and Noah D. Goodman. (2024). *Psychometric alignment: Capturing human knowledge distributions via language models*. <https://arxiv.org/abs/2407.15645>
- Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1352-1359). AAAI Press. <https://doi.org/10.1609/aaai.v31i1.10740>
- Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*, 47(2), 170-181. <https://doi.org/10.1080/1743727X.2023.2242273>
- Kokalj, E. (2026, May 22). *Gemini 3.5 Flash for vision: Evaluation and benchmarks*. Roboflow. https://blog.roboflow.com/use-gemini-3-5-flash-vision/?utm_source=chatgpt.com
- Kolesnikova, D., Fedyanin, K., Hofman, A. D., Brinkhuis, M. J. S., & Bolsinova, M. (2026). *Estimating item difficulty with large language models as experts*. arXiv. <https://doi.org/10.48550/arXiv.2605.18562>
- Li, L., & Gushta, M. (2026a, April 8-11). *Using AI judges' paired comparisons to estimate item difficulty* [Paper presentation]. National Council on Measurement in Education (NCME) Annual Meeting, Los Angeles, CA, United States.

- Li, L., & Gushta, M. (2026b). *Using content experts' paired comparison to estimate item Difficulty: What affects the accuracy?* [Manuscript under review]. Department of Research and Development, MetaMetrics.
- Linacre, J. M. (2023a). *Facets computer program for many-facet Rasch measurement* (Version 3.86.0) [Computer software]. Winsteps.com.
- Linacre, J. M. (2023b). *Facets (64-bit) help for Many-Facets Rasch measurement*. <https://www.winsteps.com/facetman64/index.htm?pairedcomparisons.htm>
- Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., & Collier, N. (2024). *Aligning with human judgement: The role of pairwise preference in large language model evaluators*. arXiv. <http://arxiv.org/abs/2403.16950>
- Liusie, A., Manakul, P., & Gales, M. (2024). LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 139–151). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.eacl-long.8>
- Lorge, I., & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement*, 12(4), 554–561. <https://doi.org/10.1177/001316445201200402>
- Loya, M., Sinha, D., & Futrell, R. (2023). Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3711–3716). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.241>
- 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>
- MetaMetrics. (2022). *Quantile Framework for Mathematics: Development and validity evidence*. https://partnerhelp.metametricsinc.com/image/Quantile%20Framework%20for%20Mathematics%20Validity%20Evidence_2022.pdf
- Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., & Bendersky, M. (2023). *Large language models are effective text rankers with pairwise ranking prompting*. arXiv. <https://arxiv.org/abs/2306.17563>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danish Institute for Educational Research (reprint 1980, expanded edition, with foreword and afterword by B. D. Wright). Chicago: The University of Chicago Press.
- Razavi, P., & Powers, S. (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms*. arXiv. <https://arxiv.org/abs/2504.08804>
- Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1(3–4), 185–216. <https://doi.org/10.1080/15305058.2001.9669470>
- Thurstone, L. L. (1927). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4), 384–400. <https://doi.org/10.1037/h0065439>
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). *Large language models are not fair evaluators*. arXiv.Org. <https://arxiv.org/abs/2305.17926>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 46595–46623). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b3e90d03fdf04da1011882-Abstract-Conference.html

A Appendix: LLM Prompts

System prompt (Persona): “You are a content expert in K-12 mathematics, specializing in item development and difficulty evaluation for math assessments.”

Simple user prompt: “You are a content expert in K-12 mathematics and a math teacher in the United States. You specialize in developing math test items. You will be provided two mathematics items developed for Grades 5 students. They may contain text, images, or both. Compare the two items and determine which one is more difficult for students. Provide your answer using one word, that is, either 'item1' or 'item2'.”

User prompt with vague criteria: “You will be shown two multiple-choice mathematics items developed for Grade 5 students. Each item contains a stem and four options, which may have or be images. Compare the two items and determine which one is more difficult for students. Consider cognitive demand and reasoning depth, linguistic and textual features, symbolic complexity, and distractor quality of the items. Before answering, think through the comparison step by step using the criteria above. Do this reasoning internally — do not include it in your response. Provide your response using one word, that is, either 'item1' or 'item2'.”

User prompt with detailed criteria: “You will be shown two multiple-choice mathematics items developed for Grade 5 students. Each item contains a stem and four options, which may have or be images. Compare the two items and determine which one is more difficult for Grade 5 students. Consider the following criteria when comparing the two items.

Cognitive demand:

(1) Items requiring a sequence of operations or distinct lines of reasoning consistently increase difficulty;

(2) Problems that demand conceptual understanding, interpreting unstated information, or mathematical modeling are harder than rote calculation.

(3) Contextualized problems (word problems) add cognitive load compared to decontextualized, purely numerical equations

Linguistic and textual features:

(1) Word length and syntactic complexity are major predictors. High vocabulary complexity or long sentences force students to translate a reading task into a math task.

(2) Extraneous information (distractor sentences or extraneous numbers within a word problem) makes it harder for students to identify the correct algorithm.

Symbolic Complexity and Format:

(1) Items heavily laden with complex formulas, fractions, or algebraic notation are inherently more difficult than those using whole numbers.

(2) In early and intermediate math, certain arithmetic operations act as heavy predictors. Concepts such as borrowing, carry-over, or inversed ordering naturally increase cognitive burden

(3) The presence of graphs, diagrams, or geometry can increase or decrease difficulty. If the visual is critical to solving the problem, it adds spatial reasoning demand.

Distractor Quality:

Plausible but incorrect distractors drastically drive up difficulty. Conversely, easily eliminated or poorly crafted distractors (like implausible numbers or giveaways) make multiple-choice questions artificially easy.

Before answering, think through the comparison step by step using the criteria above. Do this reasoning internally—do not include it in your response. Provide your response using one word, that is, either 'item1' or 'item2'.”