

Exploring AI-driven Methods for Pre-calibrating Difficulty of Mathematics Items with Visual Content

Leah Walker Jinnie Choi

Savvas Learning Company
{leah.walker, jinnie.choi}@savvas.com

Abstract

Use of AI is bringing efficiencies in replacing previously laborious field-testing with scaled item parameter pre-calibration, especially in language assessment. However, gaps in research exist in replicating this accomplishment for mathematics assessment items with visual content. This study explores the methodological options of using AI in pre-calibrating multimodal mathematics items.

1 Background

Item pre-calibration, typically done through field-testing, is required for precise, targeted and comparable measurement. Recent advances in artificial intelligence (AI) and natural language processing (NLP) have prompted researchers to investigate whether new methods can automate or replace aspects of costly human field-testing.

Prior work has explored predicting item difficulty directly from linguistic and textual features, including syntactic complexity, readability indices, semantic embeddings, and transformer-based representations (AlKhuyaey et al., 2024; Benedetto et al., 2023). Recent studies proposed an alternative approach in which, rather than directly predicting item parameters, simulated response data were generated using transformer-based large language models (LLMs) and subsequently analyzed using conventional IRT calibration procedures (Feng et al., 2025; Maeda, 2025). This approach preserved important aspects of traditional field-testing workflow while leveraging AI to approximate human response behavior.

Meanwhile, existing AI-based field-testing research has focused primarily on text-dominant

reading and language assessment domains. Comparatively little attention has been given to mathematics assessments, where item meaning is often conveyed through symbolic notation, graphs, diagrams, tables, and other visual representations that are integral to the reasoning process itself.

Therefore, before using AI for pre-calibrating these items, alternative methods demand consideration. For example, different strategies for representing visual content as textual input require attention, as transformer-based LLMs are principally designed to process sequential text input. Additionally, it is important to ensure that the representational strategies do not introduce unintended semantic variations, and to evaluate whether the transformer-based model with such input will predict response probabilities and item difficulties comparable to the ground truth empirical results.

2 Aims

This study explores the methodological options for using AI in pre-calibrating multimodal mathematics items, guided by the following questions.

RQ1. What are the different strategies to represent mathematics items with visual content as input for subsequent AI processing?

RQ2. When the visual content is transformed for LLM models, how do we ensure that the representation does not introduce accidental drift?

RQ3. What are the options to generate synthetic examinee response probabilities for these items?

RQ4. How can these results be evaluated for psychometric integrity?

3 Related Work

3.1 Methods for Encoding Visual Content in Mathematics Items

Transformer Architectures: Literature suggests three primary transformer architectures for processing mathematical items with graphical representations as input. The first type of approach relies on text-based encoder-only LLMs such as BERT (Bidirectional Encoder Representations from Transformers, [Devlin et al., 2019](#)) which improves earlier systems by reading words in human sentence bidirectionally for deep learning, and RoBERTa ([Liu et al., 2019](#)) which outperforms BERT with disentangled attention mechanism, dynamic masking and larger pre-training corpora. Because these models cannot process visual information, the visual component is first converted into a textual description via image captioning or optical character recognition (OCR), which then is concatenated with the question stem and passed to the model ([Lu et al., 2024](#)).

The second type of approach utilizes vision-language models (VLMs) or large multimodal models (LMMs) such as Open AI's GPT4 Vision or Google's Bard/Gemini, which transform images via direct extraction of multi-head self-attention features from patches of the original graphical representations ([Li et al., 2023](#); [Lu et al., 2024](#)). This approach minimizes error and redundancy by bypassing the image to text translation layer.

The third type of approach is based on deep fusion encoders which processes item text, visual components, and item-level metadata through separate encoders such as DeBERTaV3 ([He et al., 2021](#)) and DINOv3 ([Siméoni et al., 2005](#)). The resulting representations are fused using attention pooling and differential weighting of the relative importance of the text versus the image, which then passed as features in the IRT models ([Maeda & Lu, 2026](#)).

Hyperparameter Tuning: Because calibrated item datasets are significantly smaller in size than standard NLP corpora that are used for model training, preventing overfitting is the primary focus of hyperparameter tuning. For example, studies explored freezing the lower N layers of BERT, extracting embeddings from the frozen model, and feeding them into external

regressors like Ridge or Lasso instead of full end-to-end model fine-tuning ([Kapoor et al., 2025](#); [Peters et al., 2025](#); [Chen & Chen, 2026](#)). Also, methods like low-rank adaptation (LoRA) and adapter layers have been utilized to adapt larger models to small psychometric datasets without destroying pre-trained language representations ([Scarlatos et al., 2025](#); [Chen et al., 2026](#); [Ormerod & Kim, 2026](#)).

Typical loss functions used in literature include root mean squared error (RMSE) and mean absolute error (MAE) for default continuous item parameters, while evaluation metrics such as accuracy, precision, recall, F1 score, area under the curve and cross-entropy were also used when continuous parameters were discretized into ordinal bins for prediction (e.g., Easy, Medium, and Hard) ([Chen & Chen, 2026](#); [Kim, Koo & Lim, 2023](#); [Kim, Kim & Jang, 2023](#)). For search strategies, K-fold cross-validation with typically 5- or 10-fold evaluated strictly at the item or passage level is the standard, preventing inflation of performance metrics due to inadvertent leakage of visual elements or text during the training phase ([Badran & Preisach, 2025](#); [Štěpánek et al., 2023](#)).

3.2 Mitigating Drift and Bias in Visual-to-Textual Conversion

Among the encoding strategies, using image captions and 'Alt-text' type of visual-to-textual conversions can cause modality bias and accidental drift in item parameters. To prevent the drift caused by representation, recent literature suggests using vision-language transformers to better approximate the actual human task of synthesizing visual information with reading comprehension, and evaluate multimodal items in their native format ([Khan, 2026](#)). For high-stakes assessments, automated transformations require human-in-the-loop cognitive audits on construct equivalence and distractor viability ([Burke, 2025](#)). In order to evaluate the impact of LLMs' potential drift and bias, such as tendency to generate systematically easy items, researchers examine the benchmark quality measures against the item characteristic curves (ICC) ([Zhou et al., 2026](#)) and how the predicted parameters would alter downstream

ability estimation for examinees (Settles et al., 2020).

3.3 Generating Examinee Responses

Instead of mathematically simulating a dataset with 0 or 1 scores, researchers are increasingly prompting LLMs to act as "proxy examinees" to directly generate synthetic test responses. Srivatsa and colleagues (2025) evaluated this paradigm by treating various LLMs as synthetic examinees on mathematics and reading comprehension assessments. The study applied IRT to the LLM-generated responses to map them onto the same ability scale as real examinee populations, finding that without specific persona prompting, LLMs perform as significantly stronger, high-performing examinees than average human examinees. Other researchers addressed this challenge and introduced a framework which forces LLMs to act as examinees of specific abilities under a ground-truth IRT model using direct preference optimization (DPO) (Scarlatos et al., 2025). Recently, Ormerod and Kim (2026) simulated examinee responses to visual mathematics items using a multimodal LLM based on Qwen 3.5 which was fine-tuned to replicate the systematic error patterns and choice probabilities of examinees conditioned on specific latent ability levels.

3.4 Evaluation of Psychometric Integrity

Statistical Comparisons: To evaluate whether the new approaches are psychometrically sound, studies have compared transformer-based predictions to those from traditional methods such as Explanatory Item Response Models (EIRM), Linear Logistic Test Model (LLTM), and machine learning baselines such as support vector machines, random forests, and penalized regression trained on traditional NLP features (Sharpnack et al., 2024). To prove that interpreting the image actually aids the model, researchers also compared multimodal VLM predictions against text-only and vision-only predictions and examined whether the visual components contribute unique predictive variance to the item parameter beyond what the text provides (Khan, 2026; Maeda & Lu, 2026).

Evaluation of Reconstructed Items: To evaluate how well these models predict

parameters or reconstruct new items matching specific targets, the literature typically examines direct parameter recovery using RMSE and MAE that measure the overall magnitude of the prediction error, and rank correlation coefficients such as Spearman's r or Pearson's r to evaluate the relative ordering of the predicted item difficulties (Kolesnikova et al., 2026; Razavi & Powers, 2025).

4 Methods

To evaluate how representation and structural abstraction of mathematics items with visual components influence item parameter prediction and response surface estimation, we developed an empirical workflow combining techniques identified in the related works and predictive modeling, aligned with our four core research questions.

For RQ1, we examine two representation strategies to convert visual mathematical content into text. 'Alt-text' is accessibility-oriented and focused on describing what is visible. 'UI-like' description is what the graph means in the context of the mathematical task. We first convert the visual content into each representation format and then convert them into dense vector embeddings using Sentence-Transformer. These embeddings are then concatenated with item-level metadata to form the input feature set.

For RQ2, to ensure that any differences in psychometric outcomes are caused by the chosen representation strategy rather than unrelated variance introduced by the LLM, we provide textual descriptions to an LLM and prompt it to generate and reconstruct multiple-choice items preserving semantic equivalence with the original assessment items. We verify if the reconstructed items maintained the correct answer key, intended mathematical relationships, and plausible distractors. Items that fail these consistency checks are excluded to prevent them from skewing the analysis.

For RQ3, we explore transformer-based text-only response prediction models using DeBERTa, which estimates response probabilities conditional on structured item representations and simulated examinee ability. We train separate models for each representation condition. This isolation allows the model to

learn the distinct response patterns associated with the level of abstraction and structure unique to each strategy.

For RQ4, psychometric fidelity is evaluated by comparing the transformer-based text-only synthetic response generation and Rasch-generated response probabilities. After reconstructing items, generating response probabilities, and generating responses, we are performing new Rasch calibrations on those responses and comparing the recovered item difficulty parameters directly against the original, pre-calibrated item parameters using RMSE and Pearson correlation to measure how closely they align.

Sample: The study utilizes 330 multiple-choice mathematics assessment items. Each item contains visual information and was previously calibrated using the Rasch model with real examinee data.

To ensure model evaluation reflects the ability of the model to generalize to unexposed items, the calibrated item bank is partitioned into mutually exclusive training and test sets (80%/20%): The training set for optimizing the parameters of the transformer-based model, and the test set for evaluating predictive performance and psychometric approximation accuracy. The split is performed at the item level to ensure that all response patterns associated with a given item are contained within a single partition, preventing information leakage across conditions.

Examinee ability values are sampled consistently across both partitions to maintain comparability of distributions along the latent trait continuum.

5 Results

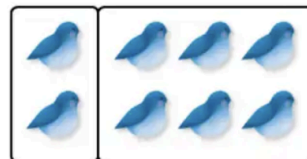
For RQ1 and RQ2, visual stimuli were converted into structured ‘UI-like’ representations (Figure 1). This strategy preserves ordinal structure while converting spatial layout into sequential linguistic form suitable for transformer-based processing.

For RQ3 and RQ4, a synthetic dataset was first constructed using a Rasch-based response function over a grid of examinee ability values. Item difficulty parameters were combined with simulated ability to generate probabilistic response values according to the Rasch model.

This item, containing a sequence of numbered squares and a missing value, was represented as: “8 identical blue birds, facing to the right as a side profile view, arranged in 2 rows of 4. The two birds in the first column are surrounded by a black rectangle. The 6 birds in the 2nd to 4th column are also surrounded by a rectangle.”

Original Item

The picture shows a way to break apart 8.
Which equation matches the picture?



- ☐ $8 = 4 + 4$
- ☐ $8 = 5 + 3$
- ☐ $8 = 2 + 6$
- ☐ $8 = 7 + 1$

Reconstructed Gemini Example

The picture shows a way to break apart 8. Which equation matches the picture?



- ☐ $8 = 4 + 4$
- ☐ $8 = 5 + 3$
- ☒ $8 = 2 + 6$
- ☐ $8 = 7 + 1$

Figure 1. Original Item and Item Reconstructed by LMM (Gemini)

Then, as preliminary baselines, ridge regression models were trained to predict Rasch-generated response probabilities using examinee ability and item embeddings across three prompt structures on a held-out test set (20%). Incorporating explicit UI-like representations achieved the strongest correlation with the response probabilities (RMSE = 0.395, $r = 0.923$).

Because full UI-like representations were not yet available for the entire item set, this baseline reflects a complete-case evaluation of a subset of items. Across the broader text representations, omitting image alt-text yielded better performance (RMSE = 0.199, $r = 0.619$) than including alt-text alongside the stem and options

(RMSE = 0.209, $r = 0.579$). Overall these preliminary findings reveal a possible tradeoff: standard textual representations provide superior absolute probability prediction, whereas explicit visual and layout descriptions capture correlations more effectively ($r = 0.923$).

6 Discussion

Results of this ongoing study provide initial evidence regarding the psychometric fidelity of methodological options when pre-calibrating mathematics items with visual content. Specifically, these preliminary findings illustrate how representational abstraction influences item parameter recovery and suggest that transformer-based text-only response models can approximate psychometric response behavior without relying on direct vision-language modeling architectures.

Initial results highlight distinct trade-offs in response probability estimation across the prompt structures. The high Pearson correlation observed under explicit UI-like representations ($r = 0.923$) demonstrates strong linear alignment between predicted and true response probabilities, even though point-estimation error was higher (RMSE = 0.395). Because full UI-like representations were evaluated on a complete-case subset due to missing layout data, ongoing work will examine whether scaling these representations across the full item bank helps reduce absolute error while preserving this high correlation.

Across the broader text representations, the preliminary finding that omitting alt-text yielded superior probability predictions (RMSE = 0.199, $r = 0.619$) compared to including alt-text (RMSE = 0.209, $r = 0.579$) points to an important feature-engineering consideration. This suggests that uncured alt-text strings may introduce semantic noise into the embedding space, whereas concise textual stems and options offer better calibrated response probability estimates. Ongoing work is exploring how visual metadata can be better structured to improve overall prediction accuracy.

Compared to traditional empirical field-testing methods, which require administering uncalibrated items to large samples of human examinees, this embedding-based pre-calibration framework offers a rapid

alternative. Traditional item response theory parameter estimation is slow, costly, and dependent on acquiring adequate sample sizes per item. By contrast, combining item embeddings with examinee ability parameters enables immediate, pre-operational response surface estimation prior to live administration. While synthetic pre-calibration is not intended to replace operational calibration entirely, it provides an efficient screening tool for automated item generation, item bank auditing, and preliminary form assembly.

A critical consideration in AI-assisted psychometric workflows is maintaining item security. Exposing operational assessment content to third-party commercial LLMs risks data leakage and model memorization, where proprietary items could be retained and exposed to examinees in generative contexts. To ensure security, our approach relies on local infrastructure execution and non-invertible feature extraction. Embedding generation is executed entirely on secure private infrastructure using open-source encoder models, preventing data transmission to external APIs or third-party training corpora. Furthermore, items and visual components are converted into fixed-dimension numerical vectors rather than processed through generative text-to-text fine-tuning. These dense vectors capture the latent semantic and structural features necessary for probability prediction while preventing the reconstruction or memorization of raw item text and answer options.

More broadly, by extending AI-based field-testing research into multimodal mathematics assessment contexts, this study contributes to emerging work at the intersection of educational measurement, NLP, and AI-driven assessment development.

References

- Samah AlKhazaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. [Text-based question difficulty prediction: A systematic review of automatic approaches](#). *International Journal of Artificial Intelligence in Education*, 34(3), pages 862–914.
- Yahya Badran and Christine Preisach. 2025. [Enhancing knowledge tracing through leakage-free and recency-aware embeddings](#). 2025. *arXiv preprint arXiv:2508.17092*.

- Lucca Benedetto, Paolo Cremonesi, Andrea Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. [A survey on recent approaches to question difficulty estimation from text](#). *ACM Computing Surveys*, 55(9), pages 1–37.
- Charles MacDonald Burke. 2025. [AI-assisted exam variant generation: A human-in-the-loop framework for automatic item creation](#). *Education Sciences*, 15(8), page 1029.
- Han Chen, Ming Li, Hong Jiao, and Tianyi Zhou. 2026. [Representing visual evidence for item difficulty prediction: Visual textualization and image-native modeling](#). *arXiv preprint arXiv:2608.04554*.
- Shi-Ting Chen and Jinsong Chen. 2026. [From text to parameters: Predicting item parameters from embedding regularization with reliability and design ceilings](#). *arXiv preprint arXiv:2607.07141*.
- Onur Demirkaya, Hsin-Ro Wei, and Evelyn Johnson. 2025. [Simulating rating scale responses with LLMs for early-stage item evaluation](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference: Full Papers*, pages 385–392.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and short papers)*, pages 4171–4186.
- Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew S. Lan. 2025. [Reasoning and sampling-augmented MCQ difficulty prediction via LLMs](#). In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 31–45. Cham: Springer Nature Switzerland.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing](#). *arXiv preprint arXiv:2111.09543*.
- 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*.
- Samin Khan. 2026. [Using vision+language models to predict item difficulty](#). *arXiv preprint arXiv:2603.04670*.
- Gun Il Kim, Sungtae Kim, and Beakcheol Jang. 2023. [Classification of mathematical test questions using machine learning on datasets of learning management system questions](#). *Plos one*, 18(10): e0286989.
- Jinsung Kim, Seonmin Koo, and Heuseok Lim. 2023. [A multi-faceted exploration incorporating question difficulty in knowledge tracing for English proficiency assessment](#). *Electronics*, 12(19), page 4171.
- Diana Kolesnikova, Kirill Fedyanin, Abe D. Hofman, Matthieu JS Brinkhuis, and Maria Bolsinova. 2026. [Estimating item difficulty with large language models as experts](#). *arXiv preprint arXiv:2605.18562*.
- Yiran Li, Junpeng Wang, Xin Dai, Liang Wang, Chin-Chia Michael Yeh, Yan Zheng, Wei Zhang, and Kwan-Liu Ma. 2023. [How does attention work in vision transformers? A visual analytics attempt](#). *IEEE Transactions on Visualization and Computer Graphics* 29(6), pages 2888–2900.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [MathVista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *International Conference on Learning Representations 2024*, pages 23439–23554.
- Hokada Maeda. 2025. [Field-testing multiple-choice questions with ai examinees: English grammar items](#). *Educational and Psychological Measurement*, 85(2):221–244.
- Hokada Maeda and Yikai Lu. 2026. [Multimodal test item parameter prediction from text, images, and metadata: Fusing together AI vision and language models](#). *Educational and Psychological Measurement*, pages 1–25.
- Christopher Ormerod and YoungKoung Kim. 2026. [Multimodal item parameter estimation using simulated response probabilities](#). *arXiv preprint arXiv:2608.10154*.
- Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou, and Robert Lissitz. 2025. [Text-based approaches to item difficulty modeling in large-scale assessments: A systematic review](#). *arXiv preprint arXiv:2509.23486*.
- Pooya Razavi and Sonya Powers. 2006. [Estimating item difficulty using large language models and tree-based machine learning algorithms](#). *International Journal of Artificial Intelligence in Education*: 100015.
- Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. [SMART: Simulated students aligned with item response theory for question difficulty prediction](#). *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25082–25105.

- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. [Machine learning-driven language assessment](#). *Transactions of the Association for Computational Linguistics*, 8, pages 247–263.
- James Sharpnack, Phoebe Mulcaire, Klinton Bicknell, Geoff LaFlair, and Kevin Yancey. 2024. [AutoIRT: Calibrating item response theory models with automated machine learning](#). *arXiv preprint arXiv:2409.08823*.
- Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. [DINOv3](#). *arXiv preprint arXiv:2508.10104*.
- KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025. [Can LLMs reliably simulate real students’ abilities in mathematics and reading comprehension?](#) In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 988–1001.
- Lubomír Štěpánek, Jana Dlouhá, and Patrícia Martinková. 2023. [Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms](#). *Mathematics*, 11(19), page 4104.
- Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang et al. 2026. [Lost in benchmarks? rethinking large language model benchmarking with item response theory](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41), pages 35085–35093.