

Simulating Examinee Responses to Figure Matrices Items Using VLMs for Item Evaluation

Onur Demirkaya and Evelyn Johnson

Riverside Insights

{onur.demirkaya, ejohnson}@riversideinsights.com

Abstract

This study evaluated whether visual embeddings from SigLIP and DINOv2 could be used to simulate responses to Figure Matrices items. Models predicted response probabilities using item screenshots and examinee ability estimates. Results showed moderate probability recovery but limited item difficulty recovery, indicating promise for early screening but not calibration replacement.

1 Introduction

Field testing is a necessary but resource-intensive stage in assessment development. Newly developed items are typically administered to human examinees before operational use so that item difficulty, discrimination, and other psychometric characteristics can be evaluated. Although this process is essential for test validity, it requires sufficient examinee samples, secure administration, and time for calibration. These challenges are especially salient for large item banks, adaptive testing programs, and assessments that require continuous development of new items.

Recent research has explored whether artificial intelligence can support early-stage item evaluation by predicting item difficulty or generating synthetic response data. Early automated item difficulty studies primarily treated item evaluation as a direct prediction problem, using item text, readability features, linguistic features, semantic embeddings, or transformer-based representations to predict item difficulty or discrimination (AlKhuyaey et al., 2023; Benedetto et al., 2023; Hsu et al., 2018; Zhou & Tao, 2020). Although these approaches are useful for preliminary item screening, they do not fully represent the test-taking process because item

difficulty is ultimately derived from examinee responses rather than item features alone.

A newer line of research has shifted from direct item difficulty prediction to AI-based response simulation. In this approach, an AI model is trained to predict response probabilities for examinees at different ability levels, and synthetic responses are generated from those probabilities. Maeda (2025) proposed an AI-examinee approach for multiple-choice English grammar items by assigning each AI examinee a latent ability value and training transformer models to predict response selection probabilities. The resulting AI-generated response data were used for psychometric analyses such as item proportion correct, item discrimination, item calibration, distractor analysis, and dimensionality analysis. Similarly, Lu and Wang (2024) used GPT-4 to create generative students with different knowledge states and compared their multiple-choice responses with real student responses. Liu et al. (2025) evaluated several large language models as synthetic respondents for college algebra items and found that LLM-generated responses could support item evaluation, although the simulated respondents showed narrower proficiency distributions than human respondents.

Extending this response-simulation framework to rating-scale assessment, Demirkaya et al. (2025) fine-tuned a DeBERTa-based model that combined item text features with examinee ability estimates to predict IRT-implied graded response probability distributions. Simulated item responses were then sampled from the predicted category probabilities and used to support psychometric analyses and early-stage item evaluation. Johnson et al. (2025) extended this line of work by comparing machine learning, text-embedding, and LLM-based approaches for generating synthetic Likert-scale

responses to social-emotional Likert scale field-test items.

These studies suggest that AI-based item evaluation may be more psychometrically meaningful when the model generates response probabilities or response data rather than only predicting a single item-level value. However, most prior response-simulation studies have focused on text-based or language-heavy items. Less is known about whether the same logic can be extended to nonverbal cognitive reasoning items, where item content is primarily visual.

Figure Matrices items provide an important setting for extending AI-based response simulation. These items require examinees to inspect a visual matrix, infer the rule or relationship among figures, and select the option that best completes the matrix. Research on Raven-like matrices suggests that item performance depends on abstract relation induction, visual encoding, rule coordination, and working-memory demands (Carpenter et al., 1990). Additional studies show that visual properties such as element salience, overlapping elements, and familiarity of shapes can influence difficulty (Meo et al., 2007; Pérez-Salas et al., 2012). Therefore, for Figure Matrices items, visual information is not incidental; it is central to both the construct and the response process.

Recent multimodal item-parameter research further motivates the use of visual models for assessment applications. Maeda and Lu (2026) proposed a multimodal framework that combines text, image, and metadata representations to predict item parameters across heterogeneous item types. Their work shows that visual features can be incorporated into item-parameter prediction, although image features may not always add unique predictive value when combined with strong text and metadata signals.

The present study applied AI-based response simulation to image-based Figure Matrices items. Rather than directly predicting item difficulty as a single item-level outcome, the model predicted the probability that an examinee with a given ability level would answer an item correctly. Synthetic dichotomous responses were then generated from these predicted probabilities. Item difficulty was evaluated after response generation by comparing item statistics derived from synthetic responses with those derived from human responses. This design parallels traditional field-testing prediction

because item difficulty emerges from response behavior.

2 Background

2.1 Figure Matrices and Visual Reasoning

Figure Matrices items are nonverbal reasoning tasks in which examinees identify visual patterns and infer relationships among elements. The items are similar in structure to Raven's Progressive Matrices, which are widely used as measures of analytic or fluid reasoning. Carpenter et al. (1990) argued that performance on matrix reasoning tasks depends on the ability to induce abstract relations and manage multiple problem-solving goals in working memory. In this view, item difficulty is related not only to the number of rules but also to the complexity of coordinating rules and elements.

Other research has emphasized the importance of visual features in matrix reasoning. Meo et al. (2007) investigated element salience and found that items were more difficult when visual elements were harder to identify, such as when elements were overlapped or unfamiliar. Pérez-Salas et al. (2012) compared contextualized and abstract matrices and concluded that differences in difficulty could be explained by element salience, verbal tagging, and item engagement. These studies suggest that visual properties of matrix items may influence examinee response behavior and should be represented in automated item evaluation models.

2.2 Visual Embeddings from Item Screenshots

Recent advances in computer vision and vision-language modeling make it possible to represent assessment item images as dense embeddings. Vision-language models (VLMs) such as SigLIP learn image representations through image-text alignment using a sigmoid loss objective (Zhai et al., 2023). Self-supervised vision models such as DINO learn image representations directly from images without requiring explicit labels and have been shown to capture transferable visual features (Caron et al., 2021; Oquab et al., 2023).

The comparison between SigLIP and DINO was included to evaluate whether vision-language alignment or self-supervised visual representation learning better supports response generation for Figure Matrices items. SigLIP may capture

visually meaningful configurations learned through image-text alignment, whereas DINO may be more sensitive to image-intrinsic structure and spatial organization. Because Figure Matrices items are nonverbal, this comparison is useful for understanding which type of pretrained visual representation is better aligned with nonverbal reasoning item behavior.

In this study, the item screenshot was used as the full visual representation of the item. The screenshot included the matrix and the multiple-choice response options as they appeared to examinees. This design choice preserved the complete test-taking display and allowed the visual encoder to represent the item as a unified visual stimulus.

3 Purpose of the Study

The purpose of this study was to evaluate whether whole-item screenshot embeddings from visual AI models could be used to generate realistic synthetic responses for image-based Figure Matrices items to support early-stage item evaluation. Specifically, the study used visual embeddings from SigLIP and DINO, together with examinee ability estimates, to predict the probability that an examinee would answer each item correctly. Synthetic dichotomous responses were then generated from the predicted probabilities and evaluated against empirical human response data. Item difficulty was evaluated after synthetic response generation, consistent with the logic of traditional field testing.

The study addressed the following research question:

RQ1. How do SigLIP- and DINO-based whole-item screenshot embeddings compare in supporting response probability prediction, synthetic response generation and item difficulty parameter recovery for Figure Matrices items?

4 Methods

4.1 Data

The study used multiple-choice Figure Matrices items from a nonverbal cognitive reasoning assessment for K-12. Items were administered by test level, with adjacent levels sharing common anchor items to support vertical linking. Each item was represented by a whole-item screenshot. The item pool included 109 unique items. Items varied

in format, including both 2×2 and 3×3 matrix designs, and the number of response options ranged from 3 to 5. Items were divided using a strict item-level split: 20 items were held out as testing items, with two testing items selected from each test level (two field test items by level), and the remaining 89 items were used as training items (see Table A1.). For the boundary levels (K & 11/12), testing items were restricted to unique items to avoid weakening the vertical linking structure. For other levels, unique items were prioritized for testing, with anchor items allowed only when necessary.

Response data sampled from K–12 students under a vertical-scale test design. We had 500 examinees at each level, 5000 examinees in total. Each examinee responded only to the items assigned to their test level; therefore, non-administered off-level items were treated as structurally missing during the concurrent calibration. A baseline Rasch calibration was conducted using the original human response data. First, a Rasch model was fit to the 89 training items to establish the item difficulty scale. These training item parameters were then fixed as anchors in a second calibration that included both training and testing items. In this fixed-anchor calibration, the 20 testing item difficulties were estimated on the same scale as the training items. Examinee ability estimates from the baseline Rasch calibration were used as person-level inputs for the response-prediction models and to compute Rasch-implied probabilities for examinee–item pairs.

4.2 Item Screenshot Processing

Only whole-item screenshots were used as image input. Each screenshot contained the item matrix and response options in the same display format as seen by examinees.

Before embedding extraction, item screenshots were resized and normalized according to the input requirements of each vision model. Screenshot preprocessing was kept consistent across training and testing items to avoid introducing artificial differences due to image handling.

4.3 Visual Embedding Models

Two pretrained visual representation models were evaluated.

The first model used SigLIP embeddings. SigLIP is a vision-language model that learns

aligned image and text representations using a sigmoid loss formulation (Zhai et al., 2023). In this study, SigLIP was used only to extract image embeddings from whole-item screenshots.

The second model used DINO embeddings. DINO is a self-supervised vision transformer approach that learns visual representations directly from image data (Caron et al., 2021; Oquab et al., 2023). DINO embeddings were included because Figure Matrices items depend heavily on visual structure, spatial organization, and abstract patterns.

For each item, the whole-item screenshot was passed through the pretrained visual encoder to obtain a fixed-length embedding. These embeddings served as item-level feature vectors in the response probability model.

4.4 Response Probability Model

The model predicted the Rasch-implied probability that examinee j would answer item i correctly. The general form of the model was:

$$\hat{P}(Y_{ij} = 1 \mid \theta_j, X_i)$$

where Y_{ij} is the dichotomous response for examinee j on item i , θ_j is the examinee's Rasch ability estimate, and X_i is the whole-item screenshot embedding for item i .

Baseline response probabilities were defined using the Rasch model. Under the Rasch model, the probability of a correct response is a logistic function of the difference between examinee ability and item difficulty:

$$P(Y_{ij} = 1 \mid \theta_j, b_i) = 1 / [1 + e^{-(\theta_j - b_i)}]$$

where b_i is the Rasch difficulty parameter for item i . These Rasch-implied probabilities served as the supervised training targets for the neural response probability model.

The neural model used a two-tower architecture to combine item information and examinee ability. In the item tower, each Figure Matrices item was represented as a whole-item screenshot. A pretrained visual encoder, either SigLIP or DINOv2, was used to extract a fixed-length image embedding from each screenshot. The resulting item embeddings were passed to a trainable feed-forward projection network. The projection network transformed the high-dimensional visual embedding into a lower-dimensional learned item representation using linear layers, nonlinear activation functions, normalization, and dropout.

This item representation was intended to capture visual information relevant to item difficulty, including the matrix structure, figures, response options, and overall visual complexity of the item.

In the ability tower, the scalar Rasch ability estimate θ_j was passed through a separate multilayer perceptron. This theta network transformed the one-dimensional ability estimate into a learned latent representation with the same dimensionality as the item representation. The purpose of this tower was to allow the model to learn nonlinear transformations of the ability scale rather than using theta only as a raw scalar input.

After the item and ability towers produced their respective representations, the model combined them into a joint representation. The joint representation included the item representation, the ability representation, their element-wise sum, and their element-wise product. The element-wise product served as an interaction term, allowing the model to capture how the relationship between item features and response probability changes across the ability scale. The combined representation was then passed through a prediction head consisting of additional fully connected layers and nonlinear activation functions. The final output was transformed into a probability between 0 and 1 and interpreted as the predicted probability of a correct response.

Binary cross-entropy with soft probability was used as the primary loss function:

$$L = -[p_{ij} \log(\hat{p}_{ij}) + (1 - p_{ij}) \log(1 - \hat{p}_{ij})]$$

where $\hat{p}_{ij}$ is the predicted probability of a correct response. This objective encouraged the model to emulate the Rasch response surface using item-image representations and examinee ability estimates rather than deterministic labels.

Model training was conducted on AWS SageMaker using Python and the PyTorch library (Paszke et al., 2019)

4.5 Synthetic Response Generation

After training, the model predicted a probability of correct response for each examinee-item pair in the evaluation set. Synthetic dichotomous responses were generated by sampling from a Bernoulli distribution:

$$Y_{ij} \sim \text{Bernoulli}(\hat{p}_{ij})$$

where Y_{ij} is the generated response (0 or 1). This probability-based sampling procedure was used to

mimic the stochastic nature of human response behavior. It also avoided treating the predicted probability as a deterministic classification rule.

The generated responses were assembled into a synthetic response matrix. These simulated responses were then calibrated using the same fixed-anchor Rasch procedure used for the baseline analysis. The resulting testing-item difficulty estimates were compared with the human-calibrated baseline difficulties to evaluate how well the model recovered item difficulty for unseen Figure Matrices items.

4.6 Evaluation

SigLIP- and DINO-based model performances were evaluated at both the response level and item level.

At the response level, predicted probabilities were evaluated using correlations and classification accuracy after applying a probability threshold. These metrics assessed whether the model produced accurate probabilities for individual examinee-item responses and reproduce responses similar to the human responses.

At the item level, item difficulty estimates recovered from the synthetic response matrix were compared with item difficulty estimates from the human response matrix. Evaluation metrics included bias, mean absolute error (MAE), root mean squared error (RMSE), Pearson correlation, and rank-order consistency.

5 Results

The testing items had a mean baseline difficulty of 0.13 with a standard deviation of 1.64, while the training items had a mean of -0.14 with a standard deviation of 1.65 (see Table A2). Thus, the two sets showed highly similar difficulty variability, although the testing set was slightly more difficult on average and had a narrower observed range (-2.44 to 2.86) than the training set (-4.15 to 2.98). Embedding similarity results also suggested that the testing items were visually comparable to the training items (see Table A3). For SigLIP, mean cosine similarity was similar across train–train (0.84), test–test (0.82), and test–train (0.83) item pairs. A similar pattern was observed for DINOv2, with mean similarities of 0.66, 0.65, and 0.65, respectively. These results indicate that the held-out testing items were broadly comparable to the training items in both Rasch difficulty distribution and visual embedding space, supporting the

validity of the item-level split for evaluating generalization to unseen items.

At the response-probability level, both vision-language models showed moderate correspondence between AI-estimated probabilities and Rasch-implied true probabilities. SigLIP produced stronger correlations than DINOv2 in both the training and testing sets. For SigLIP, the probability correlation was 0.70 for training items and 0.61 for held-out testing items. For DINOv2, the corresponding correlations were 0.68 and 0.53. As expected, correlations were higher for training items than for testing items, indicating some reduction in generalization accuracy for unseen items. Overall, these results suggest that both models captured meaningful response-probability signal, with SigLIP showing somewhat better generalization to held-out Figure Matrices items.

After probabilities were converted into synthetic 0/1 responses through Bernoulli sampling, agreement with the observed human response data was more limited. DINOv2 produced slightly higher classification accuracy than SigLIP, with accuracies of 0.59 for training items and 0.56 for testing items, compared with 0.53 and 0.52 for SigLIP, respectively (see Table 1). Precision, recall, and F1 values were generally close to 0.50 across conditions, indicating that the sampled synthetic responses only modestly reproduced individual observed response outcomes. This reduction relative to the probability-correlation results is expected because Bernoulli sampling introduces additional random variation and observed item responses contain individual-level noise beyond the Rasch-implied probability surface. Thus, while the models captured moderate probability-level signal, the synthetic dichotomous responses showed weaker response-level classification agreement with human data.

Model	Item Set	Accur.*	Prec.	Rec.	F1
DINOv2	Testing	0.56	0.50	0.50	0.52
	Training	0.59	0.51	0.53	0.52
SigLIP	Testing	0.52	0.50	0.50	0.47
	Training	0.53	0.50	0.50	0.47

Table 1. Response classification metrics (generated vs. human response data)

*Note: Accur. = Accuracy; Prec. = Precision; Rec. = Recall.

Model	RMSE	MAE	Bias	r	ρ
DINOV2	1.54	1.31	-0.21	0.29	0.33
SigLIP	1.47	1.28	-0.20	0.44	0.38

Table 2. Rasch b parameter recovery metrics

Rasch item difficulty recovery showed modest correspondence between the synthetic-response calibrations and the human-calibrated baseline estimates. Table 2 shows that SigLIP produced stronger recovery, with lower error and higher association with baseline difficulties (RMSE = 1.47, MAE = 1.28, Pearson $r = 0.44$, Spearman $\rho = 0.38$) than DINOV2 (RMSE = 1.54, MAE = 1.31, Pearson $r = 0.29$, Spearman $\rho = 0.33$). Both models showed negative bias, indicating that estimated difficulties were slightly lower than the baseline values on average. Figures 1-2 also indicate substantial shrinkage toward the center of the difficulty scale: very difficult items tended to be underestimated, whereas some easier items were estimated as less easy than in baseline calibration. Overall, SigLIP demonstrated somewhat better item-difficulty recovery than DINOV2, but both models showed limited precision in reproducing the full range and rank ordering of human-calibrated item difficulties.

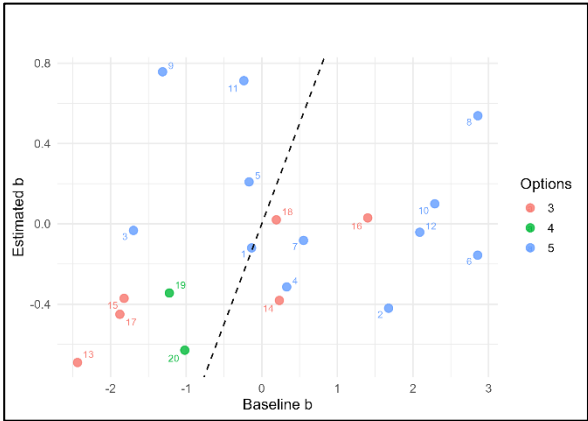


Figure 1. Estimated versus baseline (true) Rasch b parameters of testing items for DINOV2 model

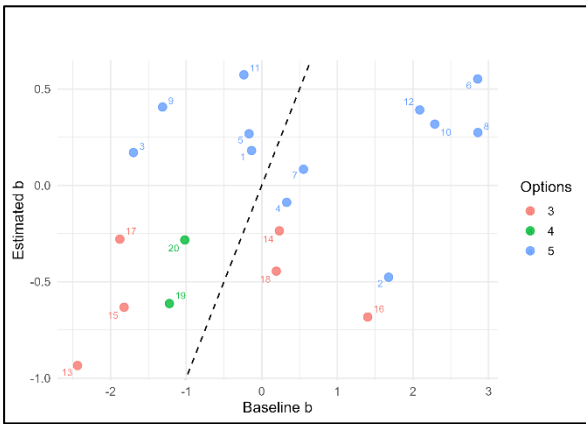


Figure 2. Estimated versus baseline (true) Rasch b parameters of testing items for SigLIP model

The ICC plots in Figure 3 compare the baseline human-calibrated curves with the SigLIP- and DINOV2-based estimated curves for the 20 held-out testing items, ordered from easier to more difficult according to the baseline Rasch difficulty estimates. For easier items, the baseline curves are shifted to the left, indicating higher probabilities of correct response at lower ability levels. For more difficult items, the baseline curves shift to the right. In particular, the AI-estimated curves tended to be shifted toward the center relative to the baseline, indicating difficulty compression: easier items were sometimes estimated as less easy, and difficult items were often estimated as less difficult. Across items, SigLIP and DINOV2 produced broadly similar ICC shapes, although SigLIP appeared frequently slightly closer to the baseline curves, consistent with its stronger item difficulty recovery metrics.

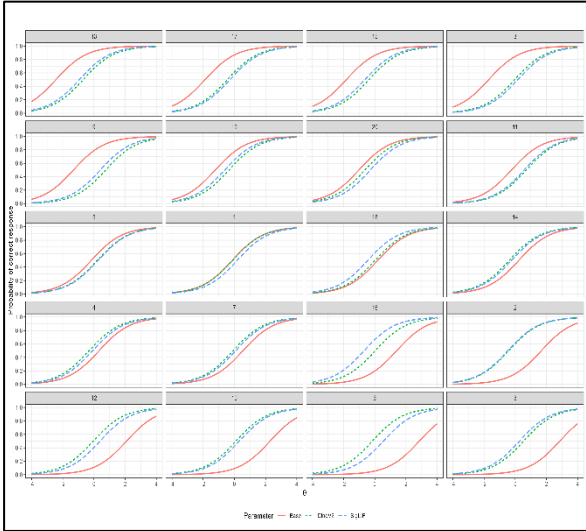


Figure 3. ICCs across testing items constructed based on true and AI-estimated difficulty values

6 Discussion and Conclusion

This study extended AI-based response simulation to image-based Figure Matrices items. Rather than directly estimating item difficulty as a single item-level target, the model predicted examinee-item response probabilities using whole-item screenshot embeddings and examinee ability estimates. Synthetic responses were then sampled from predicted probabilities and used to recover item-level statistics. The study was designed to align with the logic of field testing. This approach is particularly important for nonverbal reasoning items because difficulty may depend on complex visual features that are difficult to summarize manually.

The results suggest that pretrained visual embeddings captured meaningful but incomplete response-probability information. Both SigLIP and DINOv2 showed moderate correlations with Rasch-implied probabilities, with SigLIP generalizing better to held-out items. However, agreement weakened after probabilities were sampled into synthetic 0/1 responses, reflecting both Bernoulli sampling variability and individual-level response noise.

At the item level, Rasch difficulty recovery was modest. SigLIP outperformed DINOv2, but both models showed difficulty compression: difficult items were often underestimated, and easier items were sometimes estimated as less easy than in the human-calibrated baseline. Thus, the models captured broad probability trends but did not fully recover the range or rank ordering of item difficulty.

These findings suggest that whole-item screenshot embeddings alone may be insufficient for fully representing the cognitive and visual features that drive Figure Matrices difficulty. Matrix reasoning items often depend on abstract rule induction, relational integration, distractor plausibility, visual salience, and working-memory demands. A whole-screenshot embedding may capture general visual layout and surface similarity but may not isolate the specific rule structures or distractor relationships that influence human item difficulty.

Future work should evaluate whether adding structured item features improves performance. Features derived from the literature on matrix reasoning, such as rule count, transformation type,

relational complexity, visual density, perceptual salience, distractor similarity, and number of plausible distractors, could be combined with visual embeddings in the prediction model. Such features may help the model distinguish items that appear visually similar but differ in cognitive demand.

Another promising direction is to move beyond frozen visual embeddings. In the present study, SigLIP and DINOv2 were used primarily as pretrained feature extractors. Training or fine-tuning the visual encoder through backpropagation may allow the model to adapt its representations more directly to assessment-specific features. Full fine-tuning may be computationally expensive and may overfit when the number of unique items is small, but partial fine-tuning, adapter-based tuning, or LoRA-style approaches could provide a practical compromise.

This study has several limitations. First, AI-generated responses are not a replacement for operational field testing with human examinees. Human response data remains necessary for final calibration and validity evidence. Second, the study used a relatively small number of unique items, which limits the ability of neural models to learn robust item-difficulty patterns. Third, visual embedding models may learn superficial visual cues such as layout, density, contrast, or formatting rather than construct-relevant reasoning features. Fourth, the models were trained to approximate Rasch-implied probabilities rather than observed binary responses directly. This design supports Rasch-surface emulation but may not fully capture human response behavior. Finally, probability-based sampling introduces Monte Carlo variation; future work should consider multiple generated response matrices per item set to evaluate the stability of recovered item statistics.

Overall, the results provide cautious support for using visual AI models as early-stage tools for nonverbal item evaluation. SigLIP and DINOv2 embeddings captured meaningful response-probability signal, but item difficulty recovery remained limited. The findings suggest that whole-item screenshot embeddings may be useful for preliminary screening, identifying potentially problematic items, or prioritizing items for expert review, but they are not yet sufficient to replace empirical field testing.

References

- AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2023). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-023-00362-1>
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1-37.
- Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9650-9660).
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97(3), 404-431. <https://doi.org/10.1037/0033-295X.97.3.404>
- Demirkaya, O., Wei, H.-R., & Johnson, E. (2025). Simulating rating scale responses with LLMs for early-stage item evaluation. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 385–392). National Council on Measurement in Education. <https://aclanthology.org/2025.aimecon-main.41/>
- Hsu, F. Y., Lee, H. M., Chang, T. H., & Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6), 969-984. <https://doi.org/10.1016/j.ipm.2018.06.007>
- Johnson, E., Wei, H.-R., Wu, T., & Liu, H. (2025). Predicting and evaluating item responses using machine learning, text embeddings, and LLMs. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress* (pp. 66–70). National Council on Measurement in Education. <https://aclanthology.org/2025.aimecon-wip.8/>
- Liu, Y., Bhandari, S., & Pardos, Z. A. (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>
- Lu, X., & Wang, X. (2024). Generative students: Using LLM-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. <https://doi.org/10.1145/3657604.3662031>
- 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>
- Maeda, H., & Lu, Y. (2026). Multimodal test item parameter prediction from text, images, and metadata: Fusing together AI vision and language models [Preprint].
- Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence*, 35(4), 359-368. <https://doi.org/10.1016/j.intell.2006.10.001>
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P. Y., Li, S. W., Misra, I., Rabbat, M., Sharma, V., et al. (2023). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32. <https://arxiv.org/abs/1912.01703>
- Pérez-Salas, C. P., Streiner, D. L., & Roberts, M. J. (2012). A comparison between element salience versus context as item difficulty factors in Raven's Matrices. *Intelligence*, 40(4), 325-332. <https://doi.org/10.1016/j.intell.2012.04.001>
- Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Zhou, Y., & Tao, C. (2020). Multi-task BERT for problem difficulty prediction. In *2020 International Conference on Communications, Information System and Computer Engineering* (pp. 213-216). IEEE.

A Appendices

B	Testing		Training		All	
	N	%	N	%	N	%
Options						
3	6	30	17	19	23	21
4	2	10	7	8	9	8
5	12	60	65	73	77	71
Total	20	18	89	82	109	100

Table A1. Item allocations across items set by number of options.

Item Set	Mean	SD	Min.	Max.
Testing	0.13	1.64	-2.44	2.86
Training	-0.14	1.65	-4.15	2.98
All	-0.09	1.64	-4.15	2.98

Table A2. Descriptive statistics of true Rasch parameters by item set

Item Set	SigLIP		DINOv2	
	Mean	SD	Mean	SD
Test-Test	0.82	0.06	0.65	0.13
Train-Train	0.84	0.05	0.66	0.12
Test-Train	0.83	0.05	0.65	0.10

Table A3. Descriptive statistics of cosine similarity index for visual embeddings