

Does Including Images Improve Multimodal Language Models' Accuracy on Item Discrimination Estimation?

Jae Jun Jong¹ · Miryeong Koo¹

¹University of Illinois Urbana-Champaign

Abstract

This study measures the effectiveness of including images to predict discrimination parameters of reading items. The results suggest that providing images to language models is beneficial. By using both image and text, Lasso regressions (Tibshirani, 1996) could explain data better and random forests (Breiman, 2001) showed higher accuracy.

1 Introduction

Large Language Models (LLMs) have been extensively used in educational research, including automatic item generation (e.g., Shin & Lee, 2023), automatic scoring (e.g., Xue et al., 2025), and item parameter estimation (e.g., Han et al., 2025). Accurately predicting item parameters is important for efficient test administration and increases the amount of information that a test can provide (Yen & Fitzpatrick, 2006). In other words, educators need to accurately estimate item parameters to provide high-quality tests to test takers (Štěpánek et al., 2023). Traditionally, gathering human responses has been used to estimate item parameters (Peters et al., 2025), which is expensive (Liu et al., 2025).

Recently, many researchers (e.g., Kapoor et al., 2026; Li et al., 2026; Raina & Gales, 2024) have used LLMs to predict parameters of reading comprehension items (Peters et al., 2025). However, very few researchers consider image-related features of the reading items. While some researchers used images, their focus was not to

predict parameters of reading comprehension items (e.g., Hoyl, 2026; Khan, 2026). Therefore, prior literature suggests that LLM-based item parameter prediction for reading comprehension items rarely use images for prediction.

However, including images can positively affect students' reading comprehension performance (Guo et al., 2020). Therefore, it would also be possible for LLMs to create a richer representation of items when they use image-related features in addition to text-related features even for reading comprehension items. Furthermore, even though discrimination parameters are also important characteristics of items (Columbia University Mailman School of Public Health, n.d.), previous research (e.g., Jain et al., 2025; Kapoor et al., 2026; Raina & Gales, 2024) mainly focused on difficulty of reading comprehension items while relatively few researchers (e.g., Chen et al., 2026) predicted discrimination parameters. Therefore, this study aims to answer the following research question:

RQ: Does using image-related features in addition to text-related features significantly improve the accuracy of estimating reading comprehension items' discrimination parameters using LLMs?

Since images within reading comprehension items can positively affect test takers' reading comprehension performance (Guo et al., 2020), we hypothesize that LLMs can provide more informative embeddings when we provide images in addition to texts, which can subsequently improve the accuracy of prediction.

This study can significantly enhance our understanding of using LLMs to predict item parameter estimation, specifically focusing on effects of including images. By expanding the knowledge related to automatic item parameter estimation, this research can significantly contribute to the educational technology field by reducing the time and cost required to estimate item parameters. In sum, this research can not only improve our understanding of technology itself but also benefit students by reducing the cost required to create assessments.

2 Literature Review

2.1 Large Language Models

LLMs are machines that can understand and generate texts (Li et al., 2025; Verma et al., 2024). LLMs use embeddings, which are vectors, to understand input texts to predict likely next words (Boykis, 2024; Verma et al., 2024). Transformer architecture (Vaswani et al., 2017) enables LLMs to focus on important parts of the input and to understand long-term relationships (Verma et al., 2024). LLMs continuously change embeddings as the layer gets deeper, and these embeddings within LLMs are used for text generation, which are called hidden states (Hassani et al., 2025). Therefore, hidden states include information about how LLMs perceive input (Duan et al., 2024). These hidden states can be directly extracted and used as inputs of classification or prediction models, which can be used to predict item parameters (e.g., Kapoor et al., 2026).

2.2 Item Parameters

To understand item parameters, it is crucial to understand Item Response Theory (IRT), a mathematical framework that explains how test takers' abilities and their test performances are related (Columbia University Mailman School of Public Health, n.d.). Under the IRT framework, items have different parameters, including but not limited to difficulty and discrimination parameters (Columbia University Mailman School of Public Health, n.d.). Difficulty parameters indicate how difficult an item is and discrimination parameters indicate how well an item distinguishes between students with similar ability (Columbia University Mailman School of Public Health, n.d.). Accurately estimating item parameters is crucial for the overall quality of the

test (Štěpánek et al., 2023). In addition to test quality, item parameters can also affect the test administration process. For instance, during computerized adaptive testing, item parameters are used to select items to maximize information gain about students' abilities (Fink et al., 2025).

Traditionally, gathering human responses has been used to estimate item parameters, a process that is expensive (Liu et al., 2025; Peters et al., 2025). Therefore, many researchers are interested in automatic item parameter estimation by using LLMs that can overcome limitations of the traditional item parameter estimation method (Peters et al., 2025). In the next section, we introduce some previous studies related to automatic item parameter estimation.

2.3 Item Parameter Estimation

Recently, researchers have been using technologies such as language models to predict parameters for reading comprehension items (Peters et al., 2025). For instance, Kapoor et al. (2026) used embeddings generated by LLMs to predict the difficulty of reading comprehension items. The authors used diverse features, such as test-related features (e.g., whether questions include bolded text), text-related features (e.g., how complex the text is), and context of the test (e.g., when and where the test was used, for which age levels). The result suggests that while using more features yielded better performance, using only embeddings converted by language models can predict item difficulty to some extent.

Han et al. (2025) also aimed to predict difficulty and discrimination parameters of English items using LLMs. They fine-tuned a language model using item texts and item features (e.g., text complexity). To measure the effectiveness of including item features in fine-tuning, the authors fine-tuned the same model without item features. They found that item texts contain useful information for predicting item difficulty. In addition, they found that including item attributes during fine-tuning can improve performance. On the other hand, Han et al. (2025) reported that baseline CatBoost (Dorogush et al., 2018) showed the best performance to predict the item discrimination parameters.

In addition to converting item texts to vectors, researchers have used other methods. For instance, Raina and Gales (2024) provided text, questions, and answers to ChatGPT (OpenAI,

2022) and used the model to directly predict difficulty of items, which range from 1 to 10. In addition, the authors provided two items to the LLM and asked the model to choose a more difficult item.

Jain et al. (2025) provided questions along with other data (i.e., answers, options, student data, and information about IRT methods) to two LLMs and asked them to directly predict the difficulty and discrimination parameters of items. The authors reported that LLMs show modest performance in item parameter estimation. The authors also reported that language models generally do not assign extreme values to item parameters.

The previous literature provides two pieces of information related to automatic item parameter estimation. First, text-based features, such as embeddings generated by LLMs with item texts, can be used to predict item parameters, but providing additional information can improve the prediction accuracy (Han et al., 2025; Kapoor et al., 2026). Second, the previous studies (Han et al., 2025; Kapoor et al., 2026; Jain et al., 2025; Raina & Gales, 2024) suggest that images were not often used when predicting parameters of reading comprehension items. However, since images can positively affect humans' reading comprehension (Guo et al., 2020), it is also possible that images of items can function as an additional piece of information that can be used to improve the performance of item parameter estimation. In other words, language models can use more information from both images and item text. Therefore, to address limitations of previous studies, we aim to measure the effectiveness of including images in item parameter estimation, specifically focusing on discrimination parameters.

3 Materials and Methods

We employed three multimodal language models, Gemma-4-12b-it, Gemma-4-E2B-it, and Gemma-4-E4B-it, developed by the Gemma Team et al. (2026)¹. We chose these three models for two reasons. First, the models accept both texts and

images as input which will allow us to compare the effects of including images in addition to text, when predicting item parameters (Gemma Team et al., 2026). If separate language and vision models are used, other factors can potentially affect prediction accuracy beyond whether images are additionally included. Another reason is that we can identify the effectiveness of model sizes when predicting item parameters by using three models in the same model family (Gemma Team, 2026).

We used 152 reading assessment items written in English and their item parameters provided by the National Assessment of Educational Progress (n.d. -a, n.d. -b, n.d. -c, n.d. -d, n.d. -e, n.d. -f, n.d. -g, n.d. -h, n.d. -i, n.d. -j, n.d. -k). To convert items (i.e., texts or texts and images) to vectors, we used the hidden state of the last token at the last hidden state, motivated by earlier studies (e.g., Kapoor et al., 2026). Two hidden states were extracted from the three Gemma models with or without images (six embeddings in total). For brevity, we use the term text embeddings or text and image embeddings for the remainder of the paper. To extract hidden states, we mainly used and adapted codes provided by Gemma Team et al. (2026), Cohen (2025), and Wolf et al. (2020)².

To examine whether including images in addition to text improves accuracy on item parameter estimation, we modeled Lasso regression (Tibshirani, 1996) and random forest (Breiman, 2001). We chose Lasso regression because sizes of hidden states of Gemma models were bigger than the number of items (Gemma team et al., 2026), therefore, it is recommended to use the regression model which can select subset of inputs (Hassan et al., 2013). We chose random forest because while Lasso regression assumes linearity, random forests accommodate nonlinear relationships (Hassan et al., 2013; Raymaekers et al., 2025). The processes of model selection and model evaluation are summarized as follows: First, we divide data into training and test data with a ratio of 8:3. While it is not required to split data for random forests (Siroky, 2009), we included this process in both models for comparison purposes. Next, we fit the models

¹ Also refer to Huggingface pages of the models at <https://huggingface.co/google/gemma-4-E2B-it>, <https://huggingface.co/google/gemma-4-E4B-it>, and <https://huggingface.co/google/gemma-4-12B-it>

² Also refer to https://huggingface.co/docs/transformers/main_classes/output

using training data. Since cross validation can prevent overfitting (Arlot & Celisse, 2010), we used 5-fold cross-validation when selecting the optimal models. Particularly, the size of penalty and the number of predictors randomly selected at each split were tuned for Lasso and random forest, respectively.

We evaluated and compared the two models' explainability and predictive accuracy using test data. Specifically, root mean squared error (RMSE) were compared across the two methods (Lasso and random forest) and the three Gemma models. The RMSE measures differences between predicted and actual discrimination parameters (Miller et al., 2024), therefore lower RMSE indicates more accurate prediction. In addition, to compare multiple Lasso regressions, we used Akaike Information Criterion (AIC; Akaike, 1973 as cited in Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). AIC shows the extent that model explains data without being too complicated (Cavanaugh & Neath, 2019). BIC is similar to AIC, but it prefers simpler model than AIC (Schwarz, 1978). While we do not discuss AIC and BIC in detail in this paper, it is important to note that lower AIC and BIC are preferred (Murari et al., 2023). See Cavanaugh and Neath (2019) and Chakrabarti and Ghosh (2011) for detailed explanations for AIC and BIC.

We replicated the model selection and evaluation processes 100 times to improve the robustness of our analyses. We used the mean of 100 values that we attained from 100 replications.

4 Results

The results suggest that including image(s) in addition to texts resulted in more informative embeddings regardless of model size for reading comprehension items (see Table 1). When Gemma-4-E2B-it was used, using image and text embeddings to build Lasso regressions and random forests resulted in more accurate predictions (i.e., lower RMSE; Miller et al., 2024). Furthermore, the Lasso regression based on image and text embeddings showed lower AIC and BIC (i.e., explains data well without being too complicated; Cavanaugh & Neath, 2019; Schwarz, 1978).

When embeddings generated by Gemma-4-E4B-it were used, AIC of Lasso regression was lower when image and text embeddings were used.

However, RMSE did not decrease when compared to Lasso models built based on text embeddings. Furthermore, BIC was lower when image(s) was not used. When random forest was used, using image(s) led to more accurate predictions.

When embeddings generated by Gemma-4-12B-it were used, Lasso regression showed lower RMSE when image and text embeddings were used. In addition, both AIC and BIC of Lasso regression were lower when image and text embeddings were used. When random forest was used, using image and text embeddings led to more accurate predictions.

The results also suggest that increase in model size did not always lead to better performance. For instance, Lasso regression based on image and text embeddings generated by Gemma4-E4B-it had lower RMSE compared to that based on Gemma4-12B-it. Similarly, the AIC of the Lasso model based on image and text based embeddings generated by Gemma4-E4B-it was lower than that generated by Gemma-12B-it. However, when comparing Gemma4-E2B-it and Gemma4-E4B-it, all metrics improved for all models regardless of whether the image was used or not.

Overall, the results suggest that language models can generally generate more informative hidden states that can improve accuracy of predicting discrimination indexes, even if the items are mainly focused on text (i.e., reading comprehension items). However, increasing the size of language models did not always lead to better performance.

5 Discussion

Overall, the results suggest that including images can generally, but not always, allow language models to create more informative hidden states, which supports the hypothesis. This study also suggests that providing images is beneficial even for relatively smaller language models, such as Gemma-4-E2B-it which only has 2.3 billion effective parameters (Gemma team et al., 2026). In other words, even relatively smaller language models have the ability to use information from images to create more informative hidden states. Since many researchers in educational field have

	E2B		E4B		12B	
	Image + Text	Text	Image + Text	Text	Image + Text	Text
Lasso_RMSE	0.41	0.43	0.38	0.38	0.39	0.40
Lasso_AIC	-246.13	-226.22	-257.43	-248.82	-246.93	-240.62
Lasso_BIC	-144.59	-125.74	-147.95	-150.91	-186.12	-160.35
random forest_RMSE	0.39	0.40	0.36	0.37	0.37	0.38

Table 1: Evaluation results of six models

relatively smaller computational resources (Ormerod & Kwako, 2024), this method can be practically useful to those who are interested in estimating item parameters with relatively smaller models. Furthermore, the results (i.e., RMSE of Lasso regression models and random forests) suggest that reducing the dimensionality of hidden states to train prediction models can negatively affect the performance. This indicates reducing dimensions of hidden states results in losing information that can be used to predict item discrimination parameters. It is possible that overall hidden states include important information that cannot be used when dimensions are reduced. The results suggest that it is recommended to use models that can use entire hidden states as input, such as random forest (Salman et al., 2024) to obtain high accuracy in predicting item parameters.

In addition, the overall results suggest that there is some improvement when we use Gemma-4-E4b-it compared to Gemma-4-E2B-it. All 12 metrics were better when the model size increased. However, when comparing Gemma-4-E4B-it and Gemma-4-12B-it, not all metrics improved. It is possible that when only using embeddings generated from images and texts of items, relatively small language models are sufficient.

However, there are several limitations in this study. First, we conducted experiments with only one model family and reading items provided by National Assessment of Educational Progress. Therefore, the results may change if different models or datasets are used. For instance, items developed by individual states or schools may

have information that are more directly related to their specific context. Items that have more direct contexts may show different results compared to the study. Similarly, the items that exclusively target students whose native language is not English can also show different results from this study. Second, we mainly focused on the language models that could use both text and images as inputs (Gemma Team et al., 2026), rather than using separate vision and language models. While we assumed that using a language model that can use both images and texts would be better to identify the impact of including images, it is possible that using separate models can lead to higher accuracy.

This study introduces further possible research topics. First, as previously mentioned, it is possible to replicate the method used in this study with other datasets to investigate generalizability. Different types of datasets may have different properties, which can show different results from this study. Second, it is possible use separate models to process text and images.

Acknowledgments

The analysis was conducted using Python in Google Colab platform.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics*

- Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>
- Boykis, V. (2024). What are embeddings? https://raw.githubusercontent.com/veekaybee/what_are_embeddings/main/embeddings.pdf
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3), e1460. <https://doi.org/10.1002/wics.1460>
- Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC, and recent advances in model selection. *Handbook of the Philosophy of Science*, 7, 583–605. <https://doi.org/10.1016/B978-0-444-51862-0.50018-6>
- Chen, H., Li, M., Wang, C., Liang, Y., Zhou, D., & Zhou, T. (2026). LLMs struggle to measure what distinguishes students of different proficiency levels: A study of item discrimination in reading comprehension assessment. arXiv preprint arXiv:2606.18709. <https://doi.org/10.48550/arXiv.2606.18709>
- Cohen, M. X. (2025). LLM breakdown 4/6: Transformer outputs (hidden states). Mike X Cohen. <https://mikexcohen.substack.com/p/llm-breakdown-46-transformer-outputs>
- Columbia University Mailman School of Public Health. (n. d.). *Item Response Theory*. [https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory#:~:text=The%20scale%20for%20item%20discrimination,ranges%20between%20\(0%2C2\)](https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory#:~:text=The%20scale%20for%20item%20discrimination,ranges%20between%20(0%2C2))
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. <https://doi.org/10.48550/arXiv.1810.11363>
- Duan, H., Yang, Y., & Tam, K. Y. (2024). Do LLMs know about hallucination? An empirical investigation of LLM's hidden states. arXiv preprint arXiv:2402.09733 <https://doi.org/10.48550/arXiv.2402.09733>
- Fink, A., König, C., & Frey, A. (2025). Accounting for item calibration error in computerized adaptive testing. *Behavior Research Methods*, 57(5), 126. <https://doi.org/10.3758/s13428-025-02649-8>
- Gemma Team., Abd, S. E., Aggarwal, V., Algayres, R., Andreev, A., Bachem, O., ... & Geisler, S. (2026). *Gemma 4 technical report*. arXiv preprint arXiv:2607.02770. <https://doi.org/10.48550/arXiv.2607.02770>
- Guo, D., Zhang, S., Wright, K. L., & McTigue, E. M. (2020). Do you get the picture? A meta-analysis of the effect of graphics on reading comprehension. *AERA Open*, 6(1), 2332858420901696. <https://doi.org/10.1177/2332858420901696>
- Han, S., Rijmen, F., Boykin, A. A., & Lottridge, S. (2025). Leveraging fine-tuned large language models in item parameter prediction. In Wilson, J., Ormerod, C., & Beiting Parrish, M. (Eds.), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 250–264). <https://aclanthology.org/2025.aimecon-main.27/>
- Hassan, S. S., Farhan, M., Mangayil, R., Huttunen, H., & Aho, T. (2013). Bioprocess data mining using regularized regression and random forests. *BMC Systems Biology*, 7(Suppl 1), S5. <https://doi.org/10.1186/1752-0509-7-S1-S5>
- Hassani, H. S. (2025). *LLM Embeddings Explained: A Visual and Intuitive Guide*. Huggingface. https://huggingface.co/spaces/hesamaton/primer-llm-embedding?section=what_are_embeddings?
- Hoyle, M. (2026). Synthetic student responses: LLM-extracted features for IRT difficulty parameter estimation. arXiv preprint arXiv:2602.00034. <https://doi.org/10.48550/arXiv.2602.00034>
- Jain, Y., Hollander, J., He, A., Tang, S., Zhang, L., Sabatini, J. (2025). Exploring the potential of large language models for estimating the reading comprehension question difficulty. In Sottolare, R.A., Schwarz, J. (Eds.) *Adaptive Instructional Systems. HCII 2025. Lecture Notes in Computer Science*, vol 15812. Springer, Cham. https://doi.org/10.1007/978-3-031-92967-0_14
- Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. (2026). Prediction of item difficulty for reading comprehension items by creation of annotated item repository. arXiv preprint arXiv:2502.20663. <https://doi.org/10.48550/arXiv.2502.20663>
- Khan, S. (2026). Using vision+ language models to predict item difficulty. arXiv preprint arXiv:2603.04670. <https://doi.org/10.48550/arXiv.2603.04670>
- Li, M., Han, C., Xiao, Y., Chen, J., Jiao, H., & Zhou, T. (2026). Can LLMs estimate student struggles? Human-AI difficulty alignment with proficiency simulation for item difficulty prediction. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 25414–25441).
- Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*,

- 85(6), 1065–1090.
<https://doi.org/10.1177/0013164425134>
- 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>
- Miller, C., Portlock, T., Nyaga, D. M., & O’Sullivan, J. M. (2024). A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics*, 4, 1457619.
<https://doi.org/10.3389/fbinf.2024.1457619>.
- Murari, A., Rossi, R., Spolladore, L., Lungaroni, M., Gaudio, P., & Gelfusa, M. (2023). A practical utility-based but objective approach to model selection for regression in scientific applications. *Artificial Intelligence Review*, 56(Suppl 2), 2825–2859. <https://doi.org/10.1007/s10462-023-10591-4>
- National Assessment of Educational Progress (n.d.-a). *Reading Assessment IRT Parameters*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_irt_read.aspx
- National Assessment of Educational Progress (n.d.-b). 2022 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-c). 2017 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-d). 2013 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-e). 2011 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-f). 2009 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-g). 2007 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-h). 2005 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-i). 2003 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-j). 2002 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- National Assessment of Educational Progress (n.d.-k). 2000 Reading Assessment. The Nation’s Report Card.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- OpenAI (2022). Introducing ChatGPT.
<https://openai.com/index/chatgpt/>
- Ormerod, C. M., & Kwako, A. (2024). *Automated Text Scoring in the Age of Generative AI for the GPU-poor*. arXiv preprint arXiv:2407.01873.
<https://doi.org/10.48550/arXiv.2407.01873>
- Peters, S., Zhang, N., Jiao, H., Li, M., Zhou, T., & Lissitz, R. (2025). *Text-based approaches to item difficulty modeling in large-scale assessments: A systematic review*. arXiv preprint arXiv:2509.23486.
<https://doi.org/10.48550/arXiv.2509.23486>
- Raina, V., & Gales, M. (2024). *Question difficulty ranking for multiple-choice reading comprehension*. arXiv preprint arXiv:2404.10704.
<https://doi.org/10.48550/arXiv.2404.10704>
- Raymaekers, J., Rousseeuw, P. J., Servotte, T., Verdonck, T., & Yao, R. (2025). *A powerful random forest featuring linear extensions (raffle)*. arXiv preprint arXiv:2502.10185.
<https://doi.org/10.48550/arXiv.2502.10185>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69–79.
<https://doi.org/10.58496/BJML/2024/007>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
<https://www.jstor.org/stable/2958889>
- Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, 27(3), 27–40.
<https://doi.org/10.64152/10125/73530>

- Siroky, D. S. (2009). Navigating random forests and related advances in algorithmic modeling. *Statistics Survey*, 3, 147–163, <https://doi.org/10.1214/07-SS033>
- Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19), 4104. <https://doi.org/10.3390/math11194104>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- Verma, I., Bhatnagar, P., & Sharan, A. (2024). Text representation: A journey from traditional vector space model to LLM. In *Special Proceedings of the 26th International Annual Conference of the SSCA 2024* (pp. 127–141).
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).
- Xue, M., Liu, Y., Xiao, X., & Wilson, M. (2025). Automatic prompt engineering for automatic scoring. *Journal of Educational Measurement*, 62(4), 559–587. <https://doi.org/10.1111/jedm.70002>
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Eds.), *Educational Measurement* (4th ed., pp. 111–153). Rowman & Littlefield Publishers.