Predicting and Evaluating Item Responses Using Machine Learning, Text Embeddings, and LLMs

Evelyn S. Johnson, Hsin-Ro Wei, Tong Wu, and Huan Liu
Riverside Insights

Abstract

This study compares machine learning, text embeddings, and large language models (LLMs) for generating synthetic responses to field test items for a social-emotional assessment that uses Likert-scale responses. Using accuracy metrics and item response theory (IRT) calibration, results show that machine learning and embeddings more closely mirror student data than LLMs. Findings highlight synthetic data's promise while underscoring the need for continued validation.

1 Introduction and Background

Developing high quality assessment items requires rigorous field testing, yet this process is time consuming and costly. Traditional calibration using item response theory (IRT) typically requires hundreds of examinees per item to estimate difficulty, discrimination, and guessing parameters with acceptable precision. This burden is particularly challenging in educational settings where Likert-type surveys are widely used. Such instruments tend to measure sensitive or hard-to-predict constructs (e.g., social-emotional skills, behavioral ratings) and must often pass district level approval for wording and focus, further slowing the process of field-testing. These constraints underscore the need for alternative strategies that can accelerate item validation without compromising psychometric rigor.

Synthetic data has emerged as a promising solution for assessment developers. By generating artificial responses that approximate the distributions and relationships found in real datasets, researchers can reduce reliance on large-scale human field trials. Psychometric research demonstrates that large language models (LLMs) can serve as synthetic respondents, producing item parameter estimates that correlate highly with those derived from human data, though often with narrower variability (Liu et al., 2025). These findings suggest that artificial respondents may augment or partially substitute for actual student responses in item development.

Although LLMs provide one pathway for generating synthetic responses, they are not the only approach under investigation. A more traditional starting point has been machine learning (ML), which relies on historical student response data to predict responses to new items. However, ML models often struggle with unseen items, since new questions cannot be calibrated until sufficient student data is available. To address this limitation, researchers have explored text-aware methods that incorporate semantic information from item content. For example, Khan et al. (2025) introduced Text-LENS, which integrates text embeddings from a transformer encoder. This approach matched baseline ML performance on known items but substantially outperformed it when predicting responses to novel items (Khan et al., 2025). Such embedding-based methods offer a middle ground, more flexible than conventional ML yet more efficient than large-scale LLM simulations.

LLM-based approaches, in contrast, provide a different kind of advantage. By simulating students across ability levels, LLMs can produce synthetic response distributions that reflect difficulty trends and distractor functioning (Benedetto et al., 2024; Shridhar et al., 2023). While not perfect substitutes, these models allow test developers to "pre-pilot" items at scale, discarding poor candidates before committing resources to costly field testing.

ML, text embedding, and LLM approaches reveal a spectrum of tradeoffs. ML methods that rely on prior response patterns may be most

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 66–70

effective when field testing items that are structurally and conceptually like those previously administered. Text embedding may provide flexibility when items target new constructs within an existing domain, enabling models to generalize without extensive retraining. LLMs, while computationally intensive, may be necessary when item pools are entirely new or when the goal is to approximate the variability and reasoning patterns of real student responses. The choice among methods may depend on the purpose of item development as well as the need to balance efficiency, fidelity, and generalizability in the context of field testing.

2 Study Purpose

The purpose of this study is to compare three approaches: ML, text embedding, and LLMs, in generating synthetic responses to 10 field test items from the Devereux Student Strengths Assessment (DESSA), a standardized self-report of social-emotional competence. We evaluate the accuracy of the synthetic responses by comparing and calibrating them with IRT to compare estimated item parameters and thresholds to those derived from actual student data.

3 Methods

3.1 Sample

The sample consists of student data (N = 3,982) from an administration of the DESSA high school student self-report form (Robitaille et al., 2025). Students responded to 40 scored items and 10 field test items using a five-point Likert scale ranging from 0 (*Never*) to 4 (*Almost Always*). The demographic information of the sample is summarized in Table 1.

3.2 Measures

The DESSA (Robitaille et al., 2025) is a 50-item, standardized, norm-referenced self-report behavior rating scale for students in 9^{th} – 12^{th} grades that yields T-scores (M=50, SD=10) that are reported into three descriptive categories, "Need for Instruction", "Typical", or "Strength". The DESSA measures six social emotional competencies, Optimistic Thinking, Self-Awareness, Social Awareness, Responsible Decision Making, Relationship Skills, and Self-Management.

Category	Group	N	%
Gender	Female	1995	50.10
	Male	1987	49.90
Grade	9^{th}	1180	29.63
	10^{th}	1139	28.60
	11^{th}	883	22.17
	12^{th}	780	19.59
Race/ Ethnicity	American Indian	49	1.23
	Asian	214	5.37
	Black	1068	26.82
	Hispanic	910	22.85
	Hawaiian	6	.15
	White	1947	48.90
	Other	159	3.99
Region	Midwest	203	5.10
	Northeast	970	24.36
	South	2764	69.41
	West	45	1.1

Table 1. Demographic Information of the Sample.

3.3 Approach

We compared three methods for generating synthetic responses to the 10-field test items from the DESSA. First, we trained a Random Forest classifier on real student responses, using stratified sampling to balance classes. The model was fit to training data (n = 3,186; ~80% of the dataset) and then used to predict synthetic responses for the test set (n = 796; ~20% of the dataset). Accuracy was tracked both at the macro level and for each item.

Next, we used embeddings derived from the item text and response options to inform predictions. These embeddings were incorporated into a predictive model that mapped semantic similarity and structural features to likely student responses. As with the ML approach, predictions were generated for the 10 field-test items, with evaluation against actual student responses.

Finally, we used FLAN-T5, an instruction-tuned transformer model, to simulate student responses. The model was prompted with DESSA item stems and Likert response options, framed as if it were a high school student completing a social-emotional self-report survey. Prompts included general instructions to reflect variability in responses rather than always producing the same option to approximate realistic distributions. In addition, we

applied Low-Rank Adaptation (LoRA) fine-tuning on the training data, using real student responses to field-test items as supervised pairs, which allowed the model to better align with the rating scale and item content.

3.4 Graded Response Model

The Graded Response Model (GRM; Samejima, 1968), was employed to analyze polytomous scored items intended to measure varying levels of a latent trait. The GRM is appropriate for items with ordered categorical response options, such as those found in Likert-type scales. The probability of endorsing a response category is calculated as the difference between cumulative logistic functions across thresholds. Item calibration and model estimation were conducted using the mirt package in R (Chalmers, 2012), facilitating a robust evaluation of item functioning and trait estimation.

To evaluate the fidelity of synthetic response data generated for field items, the Pearson correlation coefficient was computed between the student-generated and synthetic response vectors (Cohen, 1988). To evaluate and compare item parameter estimates derived from synthetic data generation methods, a free calibration was conducted using the actual response dataset. From this calibration, forty item parameters were extracted and designated as anchor items. These parameters were fixed across three separate calibration conditions. The remaining ten field-test items were calibrated independently using synthetic response data. This procedure ensured that all item parameters were aligned on a common measurement scale, allowing for valid comparisons across different synthetic methods.

4 Results

We first examined the accuracy of synthetic responses generated by each approach. At the macro level, ML achieved the highest test accuracy (.62), followed closely by text embeddings (.61), while the LLM approach showed lower performance (.55). These differences were consistent across most items, with the ML and text embedding models producing comparable results, and the LLM yielding weaker alignment with observed student responses.

At the micro level, item-specific test accuracies further illustrated these trends (see Table 2). Machine learning predictions for individual field test items ranged from .50 to .71, with higher accuracy observed for items Q15 (“respect a person’s right to have a different opinion?”) and Q35 (“make others feel welcome or included?”). Text embedding results were similar, with item accuracies ranging from .48 to .71, again showing strength on items Q15 and Q35, but lower performance on Q25 (“recognize your emotions?”), and Q50 (“have a teacher or other adult at school you can talk to?”).

LLM performance was consistently lower across items, with test accuracies clustering in the .49 to .63 range. Across all ten items, both ML and text embedding methods maintained consistent predictive performance, whereas the LLM tended to underpredict or misalign with actual student response patterns (Figure 1).

Item	ML		Text Embed		LLM	
	Train	Test	Train	Test	Train	Test
7	1.00	0.63	0.88	0.61	0.57	0.56
10	1.00	0.62	0.90	0.61	0.54	0.54
15	1.00	0.71	0.89	0.71	0.52	0.48
20	1.00	0.66	0.87	0.64	0.61	0.62
25	1.00	0.53	0.86	0.53	0.53	0.53
30	1.00	0.62	0.87	0.59	0.57	0.57
35	1.00	0.67	0.87	0.65	0.62	0.59
40	1.00	0.62	0.87	0.60	0.63	0.62
45	1.00	0.67	0.89	0.66	0.53	0.49
50	1.00	0.50	0.86	0.48	0.49	0.49

Table 2. Item level accuracies across ML, text embedding and LLM approaches.

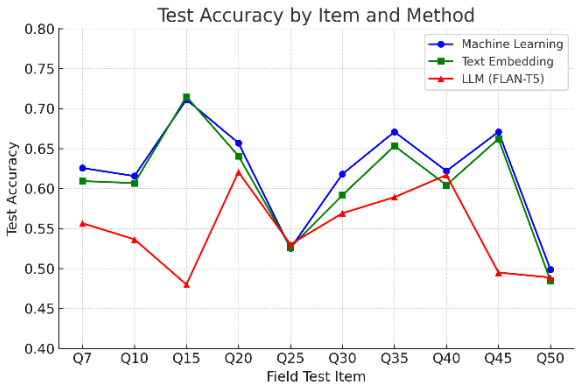

Figure 1: Accuracy levels per item across ML, text-embedding and LLM approaches.

4.1 IRT Calibration

To assess whether synthetic responses could recover psychometric characteristics of the DESSA field test items, we calibrated the 10 field items under a GRM using synthetic data from each method and compared the resulting discrimination and threshold parameters with those obtained from real student responses (Table 3).

Statistic	Method	Mean	SD	Bias	RMSE	r
a	Actual	1.86	0.36			
	ML	3.16	0.63	1.30	1.86	0.76
	Text	3.02	0.57	1.16	1.43	0.88
	LLM	5.06	0.01	3.21	10.41	-0.33
b1	Actual	-2.64	0.49			
	ML	-2.59	0.41	0.05	0.06	0.87
	Text	-2.47	0.43	0.17	0.06	0.93
	LLM	-2.32	0.03	0.32	0.33	-0.02
b2	Actual	-1.67	0.48			
	ML	-1.84	0.53	-0.17	0.05	0.95
	Text	-1.73	0.51	-0.07	0.01	0.98
	LLM	-1.65	0.01	0.02	0.21	-0.07
b3	Actual	-0.38	0.35			
	ML	-0.41	0.39	-0.03	0.00	0.99
	Text	-0.42	0.40	-0.04	0.01	0.98
	LLM	-0.30	0.01	0.08	0.12	0.36
b4	Actual	0.88	0.31			
	ML	0.87	0.30	-0.01	0.01	0.92
	Text	0.81	0.30	-0.08	0.02	0.91
	LLM	0.73	0.01	-0.15	0.10	0.48

Table 3. Estimation of Graded Response Model Item parameters (10 field items).

Across items, the ML and text embedding approaches showed broadly similar correspondence with actual parameters; neither consistently outperformed the other. By contrast, the LLM simulations exhibited weaker alignment with thresholds from actual student data and greater instability across items, echoing their lower classification accuracy. Overall while all three methods produced plausible synthetic responses, the ML and embedding approaches better preserved psychometric fidelity relative to the LLM.

5 Conclusion

This investigation highlights both the promise and the limitations of synthetic data for accelerating assessment development. Across the three synthetic data approaches applied to the DESSA field items, conventional machine learning slightly outperformed the text-embedding model, and both exceeded the LLM in aligning with observed student responses and IRT-derived item parameters. These findings indicate that ML and embeddings can plausibly support early item evaluation and calibration, while current LLM outputs appear less reliable for parameter recovery for assessments like the DESSA. Continued investigation on novel field test items aligned with different purposes (e.g., similar items for new constructs within the same domain) will inform when to use different approaches to generate synthetic data. Overall, synthetic approaches hold promise for reducing reliance on costly field testing, but continued investigation, with larger item sets, additional benchmarks, and rigorous IRT comparisons, is needed before they can be used with confidence in operational assessment.

Acknowledgments

We would like to thank our colleagues at Riverside Insights for their feedback on this work.

References

Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., & Buttery, P. (2024). *Using LLMs to simulate students' responses to exam questions*. Findings of the Association for Computational Linguistics: EMNLP 2024, 11351–11368. https://doi.org/10.18653/v1/2024.findings-emnlp.663

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Khan, A., Li, N., Shen, T., & Rafferty, A. N. (2025). *Just read the question: Enabling generalization to new assessment items with text awareness*. arXiv preprint arXiv:2507.08154. https://doi.org/10.48550/arXiv.2507.08154

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

Robitaille, J., Johnson, E.S., LeBuffe, P. A., & Naglieri, J. (2025). *Devereux Student Strengths Assessment (DESSA)–High School Edition*. Riverside Insights.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores 1. *ETS Research Bulletin Series*, *1968*(1), i-169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Shridhar, K., Bicer, H., & Liu, Q. (2023). Generating and evaluating tests for K–12 with LM item-response simulators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2210–2224). https://doi.org/10.18653/v1/2023.emnlp-main.135

A Appendices

none

B Supplementary Material

none