

Comparing θ Representation Strategies for Reconstructing Item Characteristic Curves with LLMs

Sungjin Nam

ACT Education Corp.

500 Act Dr

Iowa City, IA 52243

sungjin.nam@act.org

Abstract

We examine how strategies for representing student skill (θ) affect IRT-3PL item-parameter recovery from reconstructed item characteristic curves. We compare natural-language descriptors with signed decimal and scientific-notation anchors across anchor counts and placements. Our results show that dense numeric grids improve recovery, while non-uniform placements produce parameter-specific trade-offs, showing that θ representation matters.

1 Background

Collecting student responses for item field testing can be costly. To address this issue, previous studies have suggested using predicted item parameters to reduce calibration sample requirements (Ulitzsch et al., 2025), or recovering item parameters directly from neural text features or item metadata (Yancey et al., 2024; Kapoor et al., 2025).

Recent work has instead used generative models to predict response probabilities across student ability levels. Ormerod (2026) reframed item parameter prediction as item characteristic curve (ICC) reconstruction, predicting $p(\text{correct})$ at a set of natural-language θ anchors and deriving item parameters by fitting the IRT-2PL logistic function. Scarlatos et al. (2025) used single-digit numeric θ values to simulate open-ended responses and recover a 1PL parameter. However, these studies leave open how θ should be represented, how many anchors are needed, and whether ICC reconstruction extends to 3PL passage-based MCQs, where recovery depends on slope (a), location (b), and lower-asymptote (c) behavior.

2 Aims

This study extends LLM-based ICC reconstruction to 3PL parameters (a , b , c) on high-stakes, passage-based reading comprehension items, rather than treating θ representation as a fixed design choice.

Specifically, we ask how IRT-3PL recovery performance changes with:

- RQ1. **θ encoding types:** Descriptor, Decimal, or P10 notation.
- RQ2. **Anchor counts:** 7 to 35 θ anchors.
- RQ3. **Anchor placements:** Uniform, normal-quantile, or adaptive-local spacing.

3 Samples

The dataset contains 1,720 passage-based reading MCQs from 43 high-stakes operational forms for assessing high school students. Each form included four passages and 40 items. Each passage was about 700 words long and was linked to 10 items. Reference IRT-3PL parameters were calibrated and equated using operational responses. We used 5-fold cross-validation grouped by form.

4 Methods

4.1 Training

We fine-tuned Qwen-3-1.7B with LoRA, using rank 16, $\alpha = 128$, learning rate 5×10^{-5} , and weight decay 0.03. The baseline followed Ormerod (2026) and used 20 natural-language descriptors for θ .

Each input combined the passage, item stem, options, and a θ anchor in $[-4, 4]$. Targets were $p(\text{correct} \mid \theta)$ values from the reference 3PL ICC. The model output was the probability assigned to the correct-answer option token, and the model was trained with cross-entropy loss.

Training batches grouped multiple θ anchors per item and multiple items from the same passage, allowing the model to observe within-item ICC variation and within-passage item structure. Prompt caching reduced duplicate computation over the system prompt, passage, and item text. In this setup, the LLM acts as a curve estimator. It first

predicts response probabilities, and item parameters are recovered afterward from the reconstructed ICC.

4.2 θ Encoding types

We compared three θ encoding types. The *Descriptor* baseline used 20 ability labels from Ormerod (2026), ranging from “Critical” to “Exemplary,” which are interpretable but difficult to expand while preserving distinct meanings. *Decimal* encoding used signed numeric strings, such as +2.35, to leverage the model’s prior knowledge of numbers. *P10* notation converted values into special scientific-notation tokens for sign, digits, and exponent, such as $1.39 \rightarrow <+><1><3><9><E-2>$, following Akhauri et al. (2025). This tested whether explicit numeric tokenization offsets LLM weaknesses in handling continuous numbers (Golkar et al., 2024). Unlike descriptors, both numeric formats can add anchors mechanically without inventing new ability labels.

4.3 θ Counts

Following grid-size studies in IRT-3PL recovery (Casabianca and Lewis, 2015), we tested 7, 14, 21, 28, and 35 θ anchors. This range compares the descriptor-like budget of 20 anchors with denser numeric grids.

4.4 θ Placements

The default placement used uniform spacing over $[-4, 4]$ with single-decimal-place θ values.

At the 35-anchor setting, we also tested *normal-quantile* spacing, with 33 anchors in $[-1.89, 1.89]$, and *adaptive-local* spacing, with 14 training anchors near each item’s b plus 21 uniform anchors. Held-out adaptive-local inference used the fixed uniform grid. Both non-uniform settings used two-decimal-place θ values. These conditions tested whether allocating more anchors near the high-density or high-slope regions of the ICC improves recovery.

4.5 Parameter Recovery

For each item, predicted $p(\text{correct} \mid \theta)$ values formed a reconstructed ICC. We fit the 3PL curve by least-squares curve fitting:

$$P(\theta) = c + \frac{1 - c}{1 + \exp[-1.7a(\theta - b)]}. \quad (1)$$

This yielded estimates of (a, b, c) , which we compared with the reference parameters using fold-mean Pearson r and RMSE.

5 Results

Table 1 shows that the strongest models used numeric anchors. However, no single θ design dominated every parameter. Advantages depended on anchor count and placement.

5.1 Numeric θ Encodings Are Comparable to Language Descriptors (RQ1)

At comparable anchor counts of 20 and 21, Decimal and P10 were mixed relative to the Descriptor baseline. They were marginally better for a by RMSE and for c by Pearson r and RMSE, but worse for b by Pearson r and RMSE (Table 1). This pattern also appeared at 7 and 14 anchors. Decimal usually outperformed P10, suggesting that using familiar numeric strings was easier than learning new notation tokens in this training setting.

5.2 More θ Anchors Improve Overall Parameter Recovery (RQ2)

The Descriptor baseline was stronger for b at lower anchor counts, from 7 to 21 anchors, but numeric encodings improved as the θ grid became denser. At 28 and 35 anchors, Decimal and P10 improved recovery for nearly all 3PL parameters. Thus, numeric θ anchors may be less efficient per anchor but easier to scale, which matters when ICC fidelity depends on denser curve coverage.

5.3 Nonlinear θ Placements with Trade-offs (RQ3)

Normal-quantile placement improved a and b recovery but weakened c , consistent with concentrating anchors near $\theta = 0$ and leaving sparse tail coverage. Adaptive-local placement improved only a , indicating training-time local anchors did not transfer consistently to uniform-grid inference. The placement results suggest that central anchors help slope and difficulty, while tail coverage remains important for guessing.

6 Conclusions

In this study, we show that numeric θ encodings can match or exceed language descriptors for 3PL parameter recovery when enough anchors are used. Numeric encodings provide a useful middle ground between leveraging tokens already learned by the

pretrained language models and enabling more finely spaced θ grids. Dense numeric grids improved overall recovery, while non-uniform placement produced parameter-specific trade-offs. For this dataset, the best default is a dense Decimal grid, with normal-quantile spacing useful when a and b are prioritized over c . Future work should test larger generative models, reasoning traces, and domains beyond reading comprehension.

piloting with bert embeddings and explainable irt models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438.

References

- Yash Akhauri, Bryan Lewandowski, Cheng-Hsi Lin, Adrian N. Reyes, Grant C. Forbes, Arissa Wongpanich, Bangding Yang, Mohamed S. Abdelfattah, Sagi Perel, and Xingyou Song. 2025. Performance prediction for large systems via text-to-text regression. *arXiv preprint arXiv:2506.21718*.
- Jodi M. Casabianca and Charles Lewis. 2015. Irt item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *Journal of Educational and Behavioral Statistics*, 40(6):547–578.
- Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. 2024. xval: A continuous numerical tokenization for scientific language models. *arXiv preprint arXiv:2310.02989*.
- 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*.
- Christopher Ormerod. 2026. Reconstructing item characteristic curves using fine-tuned large language models. *arXiv preprint arXiv:2601.02580*.
- Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. Smart: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25071–25094.
- Esther Ulitzsch, Dmitry Belov, Oliver Lüdtke, and Alexander Robitzsch. 2025. Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *Journal of Educational Measurement*.
- Kevin P. Yancey, Andrew Runge, Geoffrey Laflair, and Phoebe Mulcaire. 2024. Bert-irt: Accelerating item

A Parameter Recovery and ICC Reconstruction Results

Table 1: Recovered IRT-3PL parameters by θ encoding and anchor count. Values are held-out fold means across 5-fold passage-grouped cross-validation. The best value in each column is shown in bold, and values better than the descriptor baseline are underlined. Uniform placement is evenly spaced over $[-4, 4]$; normal-quantile places 33 of 35 anchors in $[-1.95, 1.95]$; and adaptive-descent clusters training anchors around each item's reference b , with fixed uniform anchors used for held-out inference. Average columns are simple averages across α , b , and c .

Encoding	Placement	# count	$r(\alpha)$	$r(b)$	$r(c)$	Avg. r	RMSE(α)	RMSE(b)	RMSE(c)	Avg. RMSE
Decimal	uniform	7	<u>0.648</u>	0.461	<u>0.451</u>	0.487	<u>0.289</u>	0.882	<u>0.071</u>	0.414
		14	0.531	0.461	<u>0.451</u>	0.483	<u>0.270</u>	0.881	0.084	0.412
		21	0.499	0.500	<u>0.480</u>	0.496	0.288	0.825	0.065	<u>0.203</u>
		28	<u>0.644</u>	<u>0.515</u>	<u>0.480</u>	<u>0.515</u>	<u>0.280</u>	0.831	<u>0.058</u>	<u>0.203</u>
		35	<u>0.547</u>	<u>0.523</u>	0.499	<u>0.523</u>	0.270	0.812	0.068	<u>0.185</u>
	normal-quantile	35	<u>0.601</u>	0.535	<u>0.472</u>	<u>0.531</u>	0.265	0.804	0.075	0.281
		35	0.602	0.498	0.382	0.481	0.521	1.113	0.129	0.588
	adaptive-descent	7	0.533	0.459	0.429	0.454	<u>0.286</u>	0.911	0.079	0.425
		14	0.523	0.455	<u>0.450</u>	0.476	<u>0.281</u>	0.884	0.080	0.415
		21	0.507	0.513	<u>0.490</u>	0.496	0.292	<u>0.864</u>	<u>0.067</u>	<u>0.201</u>
PI0	uniform	28	0.533	<u>0.503</u>	<u>0.490</u>	0.508	0.288	0.812	0.074	<u>0.201</u>
		35	<u>0.644</u>	<u>0.507</u>	<u>0.470</u>	0.513	<u>0.272</u>	0.816	<u>0.067</u>	<u>0.205</u>
	normal-quantile	35	0.504	<u>0.410</u>	<u>0.412</u>	<u>0.401</u>	0.279	0.819	<u>0.066</u>	<u>0.200</u>
		35	<u>0.542</u>	0.478	<u>0.462</u>	<u>0.474</u>	0.503	0.953	<u>0.121</u>	<u>0.506</u>
		20	0.541	0.513	0.443	0.499	0.294	0.820	0.073	0.396
	Descriptores (Ormerod, 2026)	uniform								

Table 2: ICC reconstruction by θ encoding and anchor count. Held-out fold means across 5-fold passage-grouped cross-validation. Common-grid metrics interpolate each predicted ICC to $\theta \in [-3.75, 3.75]$ at 0.05 increments before computing item-level errors. Lower is better. Best in each column is shown in bold based on unrounded values; values better than the descriptor baseline are underlined. Rows use the same model settings as Table 2.

Encoding	Placement	θ count	ICC RMSE	ICC MAE	Normal-wtd. RMSE	Observed-grid RMSE
Decimal	uniform	7	0.101	0.084	0.129	0.094
		14	0.102	0.084	0.127	0.099
		21	0.096	0.077	0.121	0.093
		28	0.094	<u>0.077</u>	0.120	<u>0.092</u>
		35	0.092	0.075	<u>0.116</u>	0.089
	normal-quantile	35	<u>0.093</u>	<u>0.076</u>	0.116	0.114
	adaptive-local	35	0.145	0.130	0.143	0.141
	P10	uniform	7	0.104	0.086	0.132
14			0.102	0.084	0.128	0.099
21			0.095	<u>0.077</u>	0.120	<u>0.092</u>
28			<u>0.093</u>	<u>0.076</u>	0.118	<u>0.092</u>
35			0.094	<u>0.077</u>	0.120	<u>0.092</u>
normal-quantile		35	<u>0.093</u>	<u>0.077</u>	<u>0.117</u>	0.116
adaptive-local		35	0.148	0.132	0.144	0.146
Descriptors (Ormerod, 2026)		uniform	20	0.094	0.078	0.118