Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Comparing Traditional and AI-Based IRT Estimation: An Empirical Study

O.E. Apata [*†] S. T. Ajose [‡]

†Department of Educational Psychology, Texas A&M University, Harrington Tower (EDCT), TX 77843-4225, Texas, United States

‡Department of STEM Education, University of Kentucky, Lexington, 40506, Kentucky, United States

*Corresponding author. Email: apata.olukayode@gmail.com

Abstract

The surge of generative artificial intelligence tools such as ChatGPT is changing how researchers work with data and raises questions about whether AI-based tools can provide item response theory (IRT) parameters that are consistent with established maximum likelihood estimation (MLE). This study compares traditional and AI-based IRT estimation by contrasting item parameters from ChatGPT-4o with those obtained from Stata's MLE implementation. Using secondary data from 340 Indonesian vocational high school students rated on 18-character items, we fitted 1PL, 2PL, and 3PL logistic models to eight dichotomous items. ChatGPT-4o's 1PL difficulty estimates were largely consistent with Stata's and led to the same conclusion that the 2PL model fit better than the 1PL and 3PL. However, 2PL and 3PL results revealed larger discrepancies in discrimination for highly informative items and substantially higher guessing parameters under ChatGPT-4o. These findings suggest that large language models can support exploratory IRT analysis but require further validation, transparency, and methodological control before they are used for operational calibration in psychometrics.

Keywords: ChatGPT, Large language models, Item response theory, Item parameter estimation, AI in psychometrics

1. Introduction

Item response theory (IRT) has become a standard approach for educational and psychological measurement because it links item responses to an underlying latent trait and produces item parameters that can be used across samples and test forms (e.g., difficulty, discrimination, and guessing) (Huang, 2015). In practice, researchers often estimate IRT models using specialized software or general statistical packages such as Stata, which implement maximum likelihood or marginal maximum likelihood procedures for models like the one-parameter logistic (1PL, Rasch), two-parameter logistic (2PL), and three-parameter logistic (3PL) models (Baker, 1987).

At the same time, generative artificial intelligence (GenAI) tools are changing how people work with data and reshaping higher education (Author, 2025a). Large language models (LLMs) such as ChatGPT can write and run code, summarize results, and even fit complex statistical models when given appropriate prompts (Kwok et al., 2024). For measurement practitioners, this raises an important question: can AI-based tools provide item parameter estimates that are reasonably consistent with established maximum likelihood estimation (MLE) methods? If they can, AI systems could lower the technical barrier for fitting IRT models and help educators who do not have access to specialist software. If they cannot, then there is a risk that users may rely on unstable or poorly calibrated parameter estimates, especially in high-stakes testing.

To our knowledge, empirical work on AI-driven IRT estimation is still very limited. Most studies that involve AI in assessment focus on automated scoring, adaptive testing algorithms, or item generation, rather than the accuracy of item parameter estimation itself (Gardner et al., 2021). This leaves a gap in the literature

on how closely AI-generated parameters align with those obtained from traditional software and whether AI tools lead to the same model-choice decisions when comparing the 1PL, 2PL, and 3PL models.

The present study directly addresses this gap by comparing traditional and AI-based IRT estimation methods, specifically, contrasting parameter estimates derived from ChatGPT with those from Stata's MLE implementation. Our study is guided by three questions:

1. How similar are ChatGPT 4o and Stata in their item difficulty, discrimination, and guessing estimates under the 1PL, 2PL, and 3PL models?
2. Do the two approaches lead to the same conclusions about model fit when comparing the 1PL, 2PL, and 3PL models?
3. What patterns of agreement and disagreement could guide measurement practitioners who are considering AI-based IRT estimation?

2. Methods

2. 1 Dataset and items

The study used the *Dataset for 21st Century Character Exploratory Factor Analysis in Vocational High School Students* from Indonesia (Effiyanti, 2024). The analytic sample consists of $N = 340$ vocational high school students. Each student was rated by a classroom teacher who completed an 18-item Likert questionnaire (1 = strongly disagree to 5 = strongly agree) measuring character traits such as social behaviour, integrity, and information literacy.

For the present analyses, responses were first read into Stata 18 (StataCorp, 2023) and screened for data-entry errors and extreme missingness. Following common practice when moving from Likert scaling to dichotomous IRT models, we recoded each item to indicate endorsement of the positive category. Scores of 4 or 5 were coded as 1 (endorsed), and scores of 1 to 3 were coded as 0 (not endorsed).

Initial item analysis and inspection of item characteristic curves in Stata identified eight items with adequate variability and monotonic response patterns and no boundary estimates. These items (2, 3, 12, 21, 22, 26, 27, and 29 in the original questionnaire) were retained and treated as dichotomous indicators of a single latent "character" trait.

2. 2 Models (1PL, 2PL, 3PL)

Three logistic IRT models were evaluated. The one-parameter logistic (1PL or Rasch type) model assumes a single latent trait θ and estimates one difficulty parameter b_j for each item while constraining all discriminations a_j to be equal. This model treats items as equally sensitive to differences in the latent trait.

The two-parameter logistic (2PL) model relaxes this constraint by estimating an item-specific discrimination parameter a_j in addition to difficulty b_j. Higher a_j values indicate steeper item slopes and greater ability to differentiate students near the item's difficulty level.

The three-parameter logistic (3PL) model adds a lower asymptote or guessing parameter c_j, representing the probability of endorsing the item when the latent trait level is very low. All models assume a unidimensional latent trait, local independence of item responses, and a logistic response function with θ scaled to have a mean of 0 and a variance of 1.

2. 3 Estimation in Stata

Traditional analyses were carried out in Stata 18 using the irt command for item response theory models (StataCorp, 2023). The 1PL model was fitted with a common discrimination parameter, the 2PL with free discriminations and difficulties for each item, and the 3PL with a single guessing parameter shared across items to support convergence with eight items. For each model, convergence, parameter estimates, and standard errors were checked for plausibility.

Model comparison followed the usual nested likelihood ratio testing strategy. The 2PL model was compared to the 1PL model, and the 3PL model was compared to the 2PL model, using differences in log likelihood and the appropriate chi-square distribution. Item characteristic curves were exported from Stata to visualize how the eight items behaved across the latent trait continuum.

2. 4 AI estimation

To obtain AI-based estimates, we used ChatGPT-4o (OpenAI, 2024). For brevity, we refer to this model as ChatGPT in the rest of the paper. The dichotomized data set (eight items and 340 students) was exported as a comma-separated file and described to the model, including the intended 1PL, 2PL, and 3PL specifications. ChatGPT-4o was prompted to generate Python code that estimated item parameters and model fit statistics for each of the three models, using the same dichotomous items and treating θ as standard normal.

The generated scripts were run in the ChatGPT-4o environment, and the resulting difficulty, discrimination, and guessing parameters were extracted, along with the log-likelihood values for each model. These outputs were organized into tables that paralleled the Stata results, allowing direct comparison of parameter differences and likelihood ratio tests across the two estimation approaches.

3. Results

The results are organized around the three research questions.

3. 1 Similarity of item parameter estimates

For the 1PL model, ChatGPT-4o and Stata produced broadly comparable difficulty estimates. As shown in Table 1, differences in b parameters were modest, within about half a logit, with most around a third of a logit, and the rank ordering of easy and difficult items was identical across methods. These results suggest that ChatGPT can reasonably well approximate traditional Rasch-type difficulty estimates.

When discrimination and guessing are estimated (2PL and 3PL), differences become more pronounced. Table 2 shows that for Items 2 and 3, ChatGPT yielded larger discrimination parameters than Stata (e.g., 1.19 vs. 0.95 for Item 2; 0.88 vs. 0.66 for Item 3), indicating steeper item slopes. For Items 2 and 3, ChatGPT's b estimates were less negative than Stata's, implying marginally easier items. The largest discrepancies occurred in the guessing parameter: Stata's c estimates were close to zero (about .04), whereas ChatGPT's estimates were much higher (0.53 and 0.71). Thus, difficulty estimates remained relatively similar, whereas discrimination and, especially, guessing diverged more strongly.

The bar chart in Figure 1 highlights these differences in practice: the mean probability of a correct response under ChatGPT is higher than under Stata for both items. ICCs are presented for the Stata estimates only; ChatGPT-4o could not generate a comparable AI-based ICC figure. Figure 2 shows that all eight items have monotonic ICCs under the 2PL model estimated in Stata, with moderate to high discrimination and difficulties distributed across the θ continuum.

3.2 Agreement in model fit conclusions

Likelihood ratio tests comparing the 1PL, 2PL, and 3PL models are reported in Table 4. For Stata, the 2PL provided a significantly better fit than the 1PL, χ^2 (7) = 48.64, $p < .001$. ChatGPT reached the same conclusion, with $\chi^2(7) = 36.21$, $p < .001$. For the comparison between the 3PL and 2PL, both approaches suggested that moving to the 3PL did not improve fit. Stata's test was clearly nonsignificant, $\chi^2(1) = 0.29$, $p = .593$, and ChatGPT's test was also nonsignificant, $\chi^2(1) = 2.32$, $p = .128$. Thus, despite differences in individual parameter estimates, the two approaches led to the same substantive model fit decisions: the 2PL was preferred over the 1PL, and the 3PL offered little additional benefit.

3.3 Patterns of agreement and disagreement

These findings show a mixed pattern. ChatGPT aligns closely with Stata in terms of preferred model and overall item difficulty ordering. At the same time, ChatGPT produced higher discrimination for items 2 and 3, and consistently much larger guessing parameters, which yield higher predicted probabilities of success at low ability levels. For practitioners, this means that AI-based IRT estimation may be reasonable for exploratory work on difficulty and broad model selection. However, traditional MLE methods remain important when precise discrimination and guessing estimates are needed for high-stakes measurement.

4. Discussion

This study examined whether AI-based estimation using ChatGPT could replicate traditional maximum likelihood estimation (MLE) in IRT modeling. The results showed that ChatGPT's 1PL difficulty estimates were largely consistent with those from Stata, confirming that GenAI can emulate traditional psychometric reasoning when appropriately guided (Gurdil et al., 2024). However, differences became more pronounced in the 2PL and 3PL models, particularly in discrimination and guessing parameters. These discrepancies mirror findings from neural network–based estimation studies, in which AI models often yield more variable slope and asymptote parameters due to nonlinear optimization and overfitting (Zhang & Chen, 2024).

Despite these variations, both methods converged on the same model fit conclusions. ChatGPT and Stata each preferred the 2PL model over the 1PL model and found no meaningful improvement in the 3PL model. This agreement indicates that LLMs can reproduce the inferential structure underlying psychometric decision-making even if individual parameters differ slightly. Similar evidence was reported by Ruiz-Mendoza & Pedroza-Zúñiga (2025), who found that ChatGPT-generated test items achieved Rasch model fit and discrimination comparable to those of human-authored items. Collectively, these results suggest that AI-based estimation can serve as a practical exploratory tool but still requires traditional psychometric methods for final validation.

For practitioners, these findings support a careful integration of AI tools in measurement research. ChatGPT's ability to code, analyze, and explain results conversationally can make IRT more accessible for educators and researchers without advanced statistical training (Martínez-Plumed et al., 2019). However, in high-stakes contexts such as licensure testing or adaptive assessment, MLE or Bayesian estimation should remain the standard due to their theoretical transparency and precision (Reckase, 1998).

This study's limitations include a small item set, a single dataset, and the absence of control over ChatGPT's internal computational algorithms. Future research should replicate these analyses using larger, multidimensional datasets and examine how prompt structure influences parameter recovery. Incorporating transparent reporting standards and open-source validation will be essential to establishing AI-driven IRT estimation as a reliable complement to traditional approaches, consistent with broader calls for clear disclosure of AI tools, versions, and settings in scholarly work (Author, 2025b).

In summary, ChatGPT-4o demonstrates credible potential to approximate traditional IRT estimation, particularly for exploratory analysis, but full adoption in psychometrics will require continued validation, transparency, and methodological rigor.

4. Conclusion

This study offers an initial look at how AI-based IRT estimation compares with a traditional MLE implementation in Stata. For a small set of dichotomous character items, ChatGPT 4o produced difficulty estimates that were reasonably close to Stata's and led to the same conclusion that a 2PL model fit the data better than either the 1PL or 3PL. At the same time, discrimination estimates for highly informative items and especially the 3PL guessing parameters diverged substantially, suggesting that naïve AI pipelines can distort key features of the item response function. AI tools, therefore, appear useful for exploratory work or teaching, but they require further refinement, transparency, and systematic validation before being trusted for operational or high-stakes testing.

Acknowledgement

We gratefully acknowledge Tri Effiyanti for making the character education dataset publicly available. We also thank the Texas A&M University Graduate and Professional School for providing conference travel support, which made our participation at the International Meeting of the Psychometric Society possible.

References

Author (2025a). Deleted for peer review.
Author (2025b). Deleted for peer review.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*(2), 111–141. https://doi.org/10.1177/014662168701100201

Effiyanti, T. (2024). *Dataset for 21st-century character exploratory factor analysis in vocational high school students* [Data set]. Mendeley Data. https://doi.org/10.17632/8956y5n4t3.1

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Gürdil, H., Soğuksu, Y. B., Salihoglu, S., & Coşkun, F. (2024). *Integration of artificial intelligence in educational measurement: Efficacy of ChatGPT in data generation within the scope of item response theory* [Preprint]. arXiv. https://arxiv.org/abs/2402.01731

Huang, H.-Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement, 39*(5), 362–372. https://doi.org/10.1177/0146621614568112

Kwok, K. O., Huynh, T., Wei, W. I., Wong, S. Y. S., Riley, S., & Tang, A. (2024). Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and Structural Biotechnology Journal, 23*, 3254–3257. https://doi.org/10.1016/j.csbj.2024.08.006

Martínez-Plumed, F., Prudêncio, R. B. C., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence, 271*, 18–42. https://doi.org/10.1016/j.artint.2018.09.004

OpenAI. (2024a, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o

Reckase, M. D. (1998). *Item response theory: Parameter estimation techniques* [Review of the book *Item response theory: Parameter estimation techniques*, by F. B. Baker]. *Applied Psychological Measurement, 22*(1), 89–91. https://doi.org/10.1177/01466216980221009

Ruiz-Mendoza, K. K., & Pedroza-Zúñiga, L. H. (2025). Rasch-based comparison of items created with and without generative AI. *Journal of Technology and Science Education, 15*(2), 479–494. https://doi.org/10.3926/jotse.3135

StataCorp. (2023). *Stata Statistical Software: Release 18* [Computer software]. StataCorp LLC.

Zhang, L., & Chen, P. (2024). A neural network paradigm for modeling psychometric data and estimating IRT model parameters: Cross estimation network. *Behavior Research Methods, 56*, 7026–7058. https://doi.org/10.3758/s13428-024-02406-3

Appendix 1. Supplementary Materials

Table 1. Item Difficulty Estimates from the 1PL (Rasch) Model Using Stata and ChatGPT

Item	Stata (b)	ChatGPT (b)	Δb[a]
2	−1.49	−1.86	−0.36
3	−1.78	−2.26	−0.48
12	−1.51	−1.88	−0.37
21	−0.96	−1.15	−0.19
22	−0.86	−1.02	−0.16
26	−0.77	−0.90	−0.14
27	−1.20	−1.46	−0.26
29	−0.99	−1.18	−0.19

Table notes

a ChatGPT estimate minus Stata estimate (logits).

Stata estimated a common discrimination parameter under the 1PL model (a = 1.93), whereas ChatGPT fixed a to 1.00.

Table 2. Item Discrimination and Difficulty Estimates from the 2PL Model Using Stata and ChatGPT

Item	**Stata (a)**	**ChatGPT (a)**	**Δa**	**Stata (b)**	**ChatGPT (b)**	**Δb**
2	0.95	1.19	+0.24	−2.26	−1.99	+0.28
3	0.66	0.88	+0.22	−3.69	−2.96	+0.73

12	1.81	1.92	+0.12	−1.54	−1.60	−0.05
21	3.35	2.02	−1.33	−0.80	−0.86	−0.06
22	2.08	1.74	−0.35	−0.82	−0.80	+0.02
26	2.10	1.65	−0.45	−0.73	−0.71	+0.02
27	3.16	1.64	−1.52	−1.01	−1.28	−0.27
29	2.49	1.71	−0.78	−0.88	−0.97	−0.09

Table Notes
Δ = ChatGPT − Stata. Positive values indicate higher estimates from ChatGPT. Discrimination and difficulty parameters were estimated separately using the 2PL model.

Table 3. Item Parameter Estimates from the 3PL Model Using Stata and ChatGPT

Item	Stata (a)	ChatGPT (a)	Δa	Stata (b)	ChatGPT (b)	Δb	Stata (c)	ChatGPT (c)	Δc
2	1.00	1.19	+0.19	−2.12	−1.99	+0.13	0.04	0.53	+0.49
3	0.71	0.88	+0.17	−3.40	−2.96	+0.44	0.04	0.71	+0.67
12	1.87	1.92	+0.05	−1.48	−1.60	−0.12	0.04	0.27	+0.23
21	4.01	2.02	−1.99	−0.72	−0.86	−0.14	0.04	0.12	+0.08
22	2.18	1.74	−0.44	−0.76	−0.80	−0.04	0.04	0.09	+0.05
26	2.12	1.65	−0.47	−0.68	−0.71	−0.03	0.04	0.18	+0.14
27	3.20	1.64	−1.56	−0.97	−1.28	−0.31	0.04	0.32	+0.28
29	2.59	1.71	−0.88	−0.83	−0.97	−0.14	0.04	0.24	+0.19

Table Notes
Δ = ChatGPT − Stata (logits or probabilities).
Stata constrained all items to share one guessing parameter ($c \approx 0.04$); ChatGPT's values use the bottom-decile endorsement heuristic.

Table 4. Likelihood ratio tests for nested IRT models comparing Stata and ChatGPT estimates

Model comparison	χ^2 (Stata)	df (Stata)	p (Stata)	χ^2 (ChatGPT)	df (ChatGPT)	p (ChatGPT)
2PL vs 1PL	48.64	7	< .001	36.21	7	< .001
3PL vs 2PL	0.29	1	.593	2.32	1	.128

Table Notes
χ^2 values are from nested model likelihood ratio tests. Both approaches indicate improved fit from 1PL to 2PL. Neither approach shows a statistically significant improvement from 2PL to 3PL.

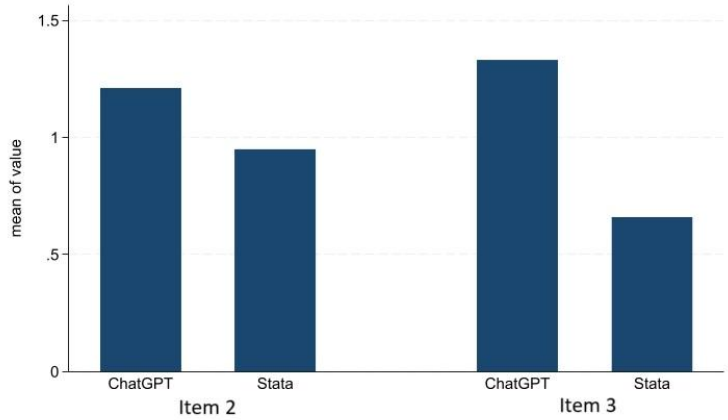

Figure 1. Comparison of mean scores for Items 2 and 3 estimated by ChatGPT and Stata.
Note. ChatGPT produced slightly higher mean values than Stata on both items.

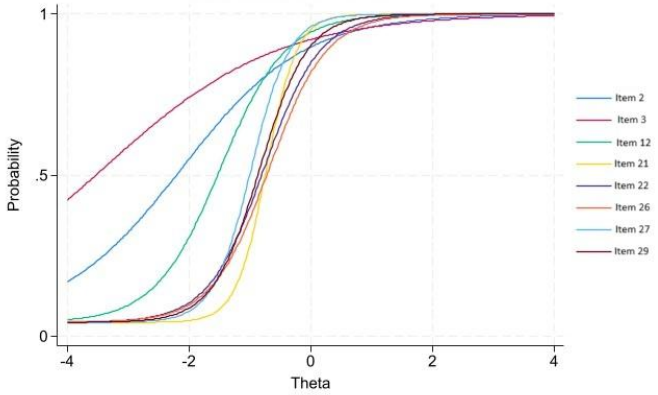

Figure 2. Item characteristic curves (ICCs) for 8 dichotomized items under the 2PL model in Stata (AI-based ICC not available)