Simulating Rating Scale Responses with LLMs for Early-Stage Item Evaluation

Onur Demirkaya, Hsin-Ro Wei and **Evelyn Johnson**
Riverside Insights
{onur.demirkaya, hsin-ro.wei, ejohnson}@riversideinsights.com

Abstract

This study explores the use of large language models to simulate human responses to Likert-scale items. A DeBERTa-base model fine-tuned with item text and examinee ability emulates a graded response model (GRM). High alignment with GRM probabilities and reasonable threshold recovery support LLMs as scalable tools for early-stage item evaluation.

1 Introduction

Field-testing is essential for developing any assessments as it serves to evaluate the statistical quality of newly developed items before the operational use. However, it remains one of the most resource-intensive and time-consuming stage in developing a test. Traditional approaches require human examinees to try out items, posing challenges related to sample availability, test security, item exposure, and scheduling . (AlKhuzaey et al., 2023; Hsu et al., 2018; Morizot et al., 2007). These challenges are growing as item banks must scale rapidly to support contemporary tests such as adaptive testing, multilingual formats, and artificial intelligent (AI)-generated content.

In response, early attempts predicted item difficulty using text-based features like syntax, semantics, word counts, embeddings, and readability indices (AlKhuzaey et al., 2023; Benedetto et al., 2023). Others used natural language processing (NLP) techniques to estimate item difficulty or discrimination (Benedetto et al., 2021; Zhou & Tao, 2020), but their accuracy remains limited. Importantly, these models often overlook distractors and fail to capture the full complexity of the human test-taking process (Benedetto et al., 2021).

More recent work has been exploring whether large language models (LLMs) can partially or fully simulate examinee responses to streamline item evaluation without sacrificing psychometric validity. For example, Lu and Wang's (2024) "generative students" prompt GPT-4 to mimic 45 learner profiles with different knowledge states, achieving a moderate correlation with undergraduate item scores ($r \approx .72$) but relying on expert-defined misconceptions and a tiny, single-topic item set. Liu, Bhandari, and Pardos (2024) go broader by blending six distinct LLMs into a 50-member ensemble, reproducing human Rasch difficulties on a small college-algebra pool ($r = .93$) yet still showing a compressed ability spread. Collectively, these studies confirm the promise of LLM-based field-testing while exposing the need for scalable methods that reduce expert overhead, widen domain coverage, and capture the full spectrum of item functioning.

Maeda (2025) moves furthest toward full AI substitution by fine-tuning 61 DeBERTa-v3 models, each tied to a specific latent ability, and embedding a two parameter logistic (2-PL) loss to predict option-level probabilities. Across 466 English-grammar items, the system matched human proportion-correct with $r = .82$ and zero mean bias, delivering plausible discrimination and distractor statistics and suggesting substantial cost and security gains. Yet achieving this required training 61 large models exposing heavy computational demand, and several extreme items still failed to calibrate accurately.

Building on Maeda's (2025) foundation, our approach leverages a single LLM that takes both item features such item and domain texts and a student's latent ability (θ) as input to predict selection probabilities of item's options, effectively emulating the graded response model

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 385–392

(GRM; Samejima, 1969). Rather than training separate models for each ability level, we condition predictions on continuous θ values and allow the model to learn item parameters implicitly from item features. This study aims to investigate if the proposed architecture enables realistic response simulation for field test Likert-scale items, supports scalable data generation, and reduces computational overhead while preserving psychometric structure, positioning it as a cost-efficient alternative for pretesting in large-scale assessments.

2 Background

2.1 Transformer Language Models

Transformer-based language models such as BERT (Devlin et al., 2018) are pre-trained on large text corpora and can be fine-tuned for various NLP tasks, including classification, summarization, and question answering. These models tokenize input text, convert it into embeddings, and process the sequence through multiple encoder layers to capture rich contextual information.

In this study, we used the DeBERTa-base model (He et al., 2021), an advanced variant of BERT and RoBERTa. DeBERTa improves representation learning by separately encoding the content and relative position of each token and computing distinct attention weights for both. This structure enhances the model's ability to capture nuanced word relationships, making it well-suited for complex language understanding tasks.

2.2 Graded Response Model

The Graded Response Model (GRM; Samejima, 1969) is a widely used item response theory (IRT) model for analyzing ordinal polytomous item responses, such as rating scales or multi-point rubrics in educational assessments. GRM models the probability that an examinee's latent ability θ_j exceeds a series of ordered thresholds for item *i*. Each item has a discrimination parameter a_i and a set of threshold (difficulty) parameters b_{ik}, one for each score category boundary. The probability of responding in category k is defined by the difference between cumulative logistic functions across adjacent thresholds:

$$P(X_{ij} = k \,|\theta_j) = P^*_{ik}(\theta_j) - P^*_{i(k+1)}(\theta_j) \quad (1)$$

where

$$P^*_{ik}(\theta_j) = \frac{1}{1 + \exp\left[-a_i(\theta_j - b_{ik})\right]} \quad (2)$$

The GRM assumes monotonicity, unidimensionality, and local independence, and it enables estimation of both person ability and item parameters on a common scale. It is especially well-suited for assessments where responses reflect degrees of correctness or agreement rather than binary outcomes. For detailed discussion, see Samejima (1969) or Baker & Kim (2004).

2.3 AI-Based Field-Test Data Generation Pipeline

This approach builds on Maeda's (2025) architecture but ease its computational cost by training a single generalized model instead of separate models per ability level and eliminating the sampling process, supporting scalable and flexible field test data generation for Likert-scale items. The model is trained on operational items with known GRM-based probabilities, using both item features and latent ability ($\boldsymbol{\theta}$) as inputs. Once trained, the model generalizes to predict probabilities for new items by conditioning on the item's text representation and examinee ability. Simulated responses are generated by sampling from the predicted probabilities, enabling psychometric analyses such as pre-calibration and item screening without requiring real student administrations. The AI-based field test data simulation pipeline is provided in Figure A1.

Process Item Features Data

We used items of the Devereux Student Strengths Assessment (DESSA). It is a standardized, strength-based behavioral rating scale developed to assess social-emotional competencies in children and adolescents (LeBuffe, et. al., 2009). The DESSA has eight empirically derived domains: self-awareness, social awareness, , self-management, goal-directed behavior, relationship skills, personal responsibility, decision making, and optimistic thinking (Shapiro & LeBuffe, 2004). Items are rated on a 5-point Likert scale (from "never" to "almost always") and yield standard scores with T-score interpretation.

Taking advantage of Likert-scale with same options across items, we only included item stem and its domain information as a text input enabling the model items as unified constructs and deeper theta-text interaction. Item stems were paired with the item's domain label (e.g., "Domain: Self-Awareness") to provide semantic context. The resulting domain-qualified text was used as input features in training the model to predict graded response probability distribution overall possible options (see Figure A2).

Calculate Conditional IRT Probabilities

To generate model training targets, we used the GRM (Samejima, 1969) to compute the conditional probability of each ordinal response category. A total of 1,000 examinee ability levels ($\boldsymbol{\theta}$) were sampled from a standard normal distribution $N(\mu = 0, \sigma^2 = 1)$, which closely approximates the ability distribution of the target population, $\mu = -0.002, \sigma^2 = 0.98$.

For each item-person pair, we used the item's GRM parameters, discrimination parameter a_i and threshold parameters b_{ik}, to compute the probability of a response in category k as given in equation 1. This yields a vector of conditional probabilities across all response categories for each item-θ pair. These probability vectors were used as target labels in training the model to emulate the GRM response function.

Fine-tune transformers with item features and theta

The fine-tuning pipeline employs DeBERTa-base as the text encoder, leveraging its disentangled-attention backbone to yield a 768-dim CLS embedding that captures both content and relative-position information efficiently (He et al., 2021). To make the single latent-ability estimate θ competitive in that high-dimensional space, the model feeds θ through a dedicated "ThetaEncoder" sub-network before concatenation. This process let the network learn either a simple or a richly nonlinear transformation as needed. It is first passed through three hidden layers (sizes 64 → 128→ 256 with GELU activations, LayerNorm, and dropout followed by tanh), producing a θ-embedding that shares scale and distributional properties with the transformer hidden states. This vector is concatenated with the original CLS embedding, giving a 1536-dim joint feature on which a dropout-regularized linear head (1536 → 5) predicts raw logits that are converted to predictive probabilities via soft-max before any loss is computed. Optimization uses cross-entropy with soft targets**:** for every training example the target distribution is the five-category probability vector produced by Samejima's graded-response IRT model, and the loss

$$CEL = -\sum_k P_k \, log\hat{P}_k \tag{3}$$

encourages the network to reproduce the entire curve rather than just the arg-max label. Because θ now enters through hundreds of weights instead of one and the loss supplies dense probabilistic feedback, the model learns item-specific category curves that vary smoothly with ability.

Generate Item Responses

Once the fine-tuned LLM model has produced a five-element probability vector $\hat{P}_{ijk} = (\hat{P}_{ij0}, \ldots, \hat{P}_{ij4})$ for examinee j on item i, to mimic human-like stochasticity probability based sampling is used to generate a concrete response rather than deterministic arg-max which can distort the latent-response surface and inflate slope estimates later in calibration. A complete response matrix is produced in this way for both training and field-test items for further psychometric analysis.

3 Methods

This study uses DESSA items to simulate a scenario where some set of previously calibrated items are available for training, while a smaller set of new items, represented only by their text, requires field-testing. Item parameters derived from prior field-testing are treated as true item parameters for both training and evaluation.

The dataset included 50 DESSA items, a standardized, strength-based behavioral rating scale developed to assess social-emotional competencies in children and adolescents. The instrument encompasses eight empirically derived domains: self-awareness, social awareness, self-management, goal-directed behavior, relationship

skills, personal responsibility, decision making, and optimistic thinking. Each item is rated on a 5-point Likert scale (from "never" to "almost always"). All items had been previously calibrated using the GRM based on responses from a nationally representative sample of 1,350 middle school students. Among the 1,350 respondents, 2.5 %, 8.4 %, 26.3 %, 35.0 %, and 27.7 % endorsed categories 0, 1, 2, 3, and 4, respectively. Overall, nearly two-thirds of the calibration sample endorse the item at a high level, while only about one in ten fall at the negative end of the scale.

To simulate the field-testing context, the items were randomly divided into 85% training items ($n = 38$) and 15% field-test items ($n = 12$), with the constraint that the field-test subset included at least one item from each SEL domain. The training items served as calibrated, operational items**,** used to fine-tune the language model and anchor the score scale during calibration. The field-test items, excluded from model training, represented new, uncalibrated items used to evaluate model generalization and calibration accuracy.

The DeBERTa-base model (He et al., 2021) was fine-tuned using the item features (domain label and item text) from the 38 training items, along with 1,000 latent ability values (θ) sampled from a standard normal distribution, $N(0,1)$, which reflects the target population's ability distribution. The AdamW optimizer (Loshchilov & Hutter, 2017) was used to minimize the CEL between the GRM-derived target probabilities and the model's predictions (James et al., 2023). Fine-tuning was conducted using the PyTorch library (Paszke et al., 2019) on a Google Colab Pro instance equipped with a NVIDIA A100 Tensor Core 40GB GPU. The model was trained for 15 epochs with a batch size of 16 per device, a learning rate of 2×10^{-5}, and a weight decay of 0.01. Item response data were generated based on $\hat{P}_{ijk}$ for all training and field-test items.

To assess the psychometric quality of the generated data, field-test items were calibrated using the GRM. (Samejima, 1969). The mean and variance of the latent ability (θ) were freely estimated, while the parameters of the training items were anchored by fixing them to their known discrimination (a_i) and threshold (b_{ik}) values, ensuring that field-test items were placed on the same scale. Calibration was conducted using the mirt package in R (Chalmers, 2012). Item parameters previously obtained from field-testing with real human examinees were treated as true values. Estimates derived from the model were evaluated against these true values using mean signed bias, mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficients (r) for each parameter.

4 Results

Table A1 shows that the average item parameters are generally comparable between the training (38 items) and testing sets (12 items). Standard deviations are also consistent across sets, with slightly more variation in the testing items' thresholds, likely due to fewer items. Overall, these similarities suggest that the item parameter distributions are almost balanced across the training and testing subsets.

Figure 1 demonstrates that the model's predicted probabilities track the true category probabilities almost perfectly in testing set (r = 0.97). The better trend observed on the training set (r = 0.99), too. This tight alignment indicates that the model captured the underlying response tendencies with high fidelity which is an essential prerequisite for downstream psychometric uses such as stochastic response simulation and item-parameter recovery.

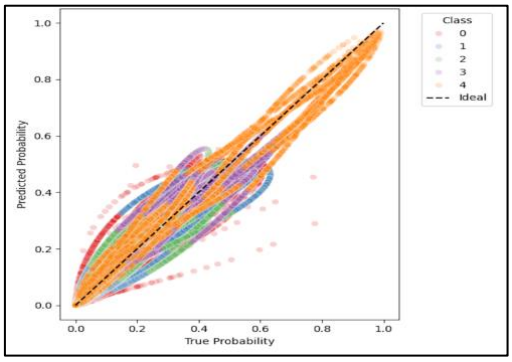

Figure 1: Predicted versus True Probabilities across Response Categories for Testing Items

However, when we translated these well-calibrated probability vectors into single categorical responses (via one draw per item to

mimic human variability rather than using deterministic arg-max sampling), discrimination among adjacent score levels became more challenging, especially for the rarer categories (0-1). Table A2 details this pattern, reporting precision, recall, and F1 for every category, together with the overall Cohen's κ that reflects agreement beyond chance. In brief, the model delivers highly calibrated probabilities and moderate-to-strong categorical accuracy where it matters most (levels 2–4), while the expected drop in metrics for the low-frequency categories reflects both class imbalance and the deliberate injection of sampling noise.

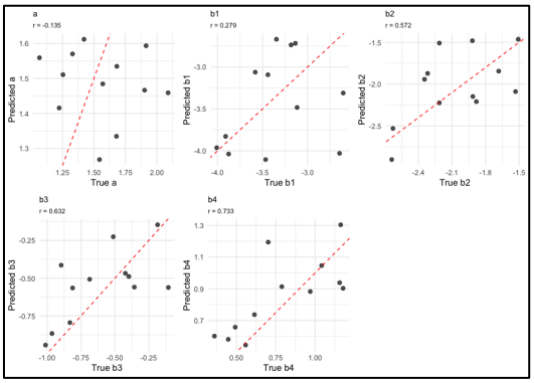

Figure 2: Scatterplots of Predicted versus True Item Parameters for Testing Items

Figure 2 showing field test items' predicted and true IRT parameters and Table 1 indicating overall numerical fit indices of those parameters together provide a consistent picture. For discrimination parameter (a), estimates were weakly and negatively correlated with the true values (r = –0.13), showing both noticeable scatter in Figure 2 and moderate error (RMSE ≈ 0.33). The slight negative bias (-0.07) and compressed range suggested the model flattens steep items and inflates shallow ones. For difficulty thresholds (b's), recovery improved monotonically from b_1 to b_4. The first threshold was the noisiest ($RMSE \approx 0.59, r = 0.28$), but accuracy doubled for the upper thresholds (b_3, b_4) where RMSE fell below 0.25 and correlations climbed above 0.60. The bias pattern was small and positive for $b_2 - b_4$, implying a slight right-shift of predicted step locations. Overall, slopes were poorly recovered, whereas later thresholds were estimated with moderate precision; early thresholds remained a concern.

Parameter	Bias	MAE	RMSE	r
a	-0.07	0.30	0.33	-0.13
b_1	-0.06	0.48	0.59	0.28
b_2	0.05	0.31	0.37	0.57
b_3	0.05	0.19	0.24	0.63
b_4	0.07	0.17	0.21	0.73

Table 1. Parameter Recovery Metrics for Testing Items

Figure 3 overlays the true (solid) and predicted (dashed) category response curves for a sample of items. For most items the ordering of curves was preserved and each predicted peak occurred near the true modal θ, confirming that the threshold structure was broadly captured. Consistent with the numeric bias, predicted curves often shift rightward, especially for the b_1 and b_2 steps, causing lower categories to dominate a wider θ-range than intended. Flattened peaks and broader overlaps reflected the underestimated discriminations, explaining why slope recovery was weak yet the model still yielded plausible probabilities.

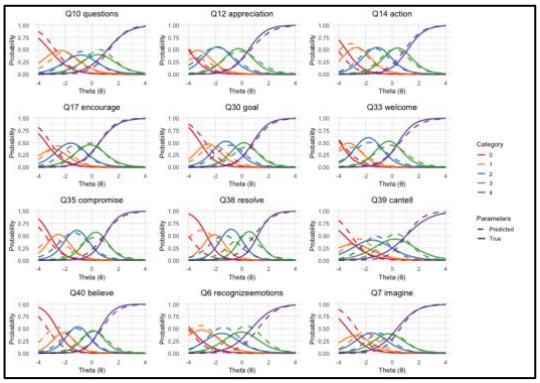

Figure 3: Predicted versus True Category Curves by Testing Items

5 Conclusions

AI-based field-testing approach in this study aims to improve the efficiency of traditional pretesting by simulating human examinee responses using AI, thereby reducing or if possible, eliminating the need for large-scale human data collection. Specifically, we investigated if the proposed approach could emulate graded response model by using a single DeBERTa-base model with item text and examinee θ to generate realistic responses to Likert-type rating scale items. The current study demonstrated that IRT statistics

derived from AI-generated responses show moderate alignment with those obtained from human examinees. This suggests that the proposed architecture can approximate key features of human response behavior in rating-scale assessments and serve as a scalable tool for early-stage item evaluation.

Item-parameter recovery paints a nuanced picture: the model captures later thresholds ($b_3 - b_4$) with reasonable precision ($RMSE \leq 0.25, r \geq 0.63$) and preserves the qualitative ordering of category response curves, yet it underestimates discrimination (a) and the earliest threshold (b_1). These findings suggest that the architecture faithfully encodes item difficulty structure but still compresses slope information, a pattern consistent with the "flattened ICC" or items with negative discriminations documented in transformer-generated response data (Byrd & Srivastava, 2022; Maeda, 2025).

This study, while promising, has several limitations that warrant consideration. First, the item pool was restricted to a small set of Likert-type social-emotional learning items, limiting the generalizability of findings to other domains. Second, although the use of stochastic sampling from predicted probabilities offers a realistic alternative to deterministic predictions, it also introduces additional variance that can inflate classification error and reduce parameter recovery precision. Future implementations should incorporate multiple draws from the predicted probability distributions to reduce Monte Carlo variance by using Rubin's Rule (1987). Third, item parameter estimation was conducted on a relatively small number of training and testing items, which may limit the robustness of recovery analyses, particularly for slope parameters. Benedetto (2023) showed that the predictive power of transformers increased with increasing training sample size; therefore, the results of the current study may increase with larger number of training items. Finally, the study relied on a single pretrained DeBERTa model; further work is needed to explore how different model architectures, sizes, and fine-tuning strategies influence response quality and psychometric fidelity.

By modeling probabilistic item responses through a single transformer-based model and evaluating their psychometric viability, this study offers a scalable pathway toward AI-enhanced pretesting workflows. While improvements are needed, particularly in recovering item discriminations, the strong probability calibration and promising threshold estimates position this approach as a compelling tool to reduce workload and improve the speed and consistency in the traditional field-testing pipelines, especially in low-resource or early development contexts.

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,* 1–53. https://doi.org/10.1007/s40593-023-00362-1

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Benedetto, L. (2023). A quantitative study of NLP approaches to question difficulty estimation (arXiv preprint arXiv:2305.10236). https://arxiv.org/abs/2305.10236

Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., & Turrin, R. (2021). On the application of transformers for estimating the difficulty of multiple-choice questions from text. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pila´n, H. Yannakoudakis & T. Zesch (Eds.), *Proceedings of the 16th workshop on innovative use of NLP for building educational applications* (pp. 147–157). Association for Computational Linguistics.

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.

Byrd, M., & Srivastava, S. (2022). Predicting difficulty and discrimination of natural language questions. In S. Muresan, P. Nakov & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics: Short papers 602 (Vol. 2,* pp. 119–130). Association for Computational Linguistics.

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

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* (arxiv preprint arxiv:1810.04805). https://arxiv.org/abs/1810.04805

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention* (arxiv preprint arxiv:2006.03654). https://arxiv.org/abs/2006.03654

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.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning*: With applications in python. Springer.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *The Devereux Student Strengths Assessment (DESSA)*. Kaplan Press.

Liu, Y., Bhandari, S., & Pardos, Z.A. (2024). Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis. *ArXiv, abs/2407.10899*.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization* (arxiv preprint arxiv:1711.05101). https://arxiv.org/abs/1711.05101

Lu, X., & Wang, X. (2024). Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*.

Maeda, H. (2025). Field-testing multiple choice questions with AI examinees: English Grammer Items. *Educational and Psychological Measurement, 85*(2), 221-244. https://doi.org/10.1177/00131644241281053

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). Guilford Press.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, *17*(4), 2. doi:10.1002/j.2333-8504.1968.tb00153.x.

Shapiro, V. B., & LeBuffe, P. A. (2004). Strength-based assessment in children: The Devereux Early Childhood Assessment and the Devereux Student Strengths Assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (2nd ed., pp. 215–236). Guilford Press.

Zhou, Y., & Tao, C. (2020). Multi-task BERT for problem difficulty prediction. *In 2020 international conference on communications, information system and computer engineering (CISCE)* (pp. 213–216). Institute of Electrical and Electronics Engineers.

A Appendices

	Parameters	Training Items	Testing Items	All Items
Mean	a	1.49	1.56	1.51
	b_1	-3.32	-3.36	-3.33
	b_2	-1.94	-2.07	-1.97
	b_3	-0.47	-0.60	-0.50
	b_4	0.96	0.79	0.92
SD	a	0.28	0.31	0.28
	b_1	0.70	0.45	0.65
	b_2	0.52	0.39	0.49
	b_3	0.40	0.31	0.38
	b_4	0.40	0.30	0.38

Table A1: Descriptive Statistics of Calibrated Item Parameters by Dataset

Data	Category	Precision	Recall	F1-Score	Kappa	*r*
Training	0	0.13	0.12	0.13	0.16	0.99
	1	0.22	0.23	0.22		
	2	0.36	0.35	0.36		
	3	0.39	0.39	0.39		
	4	0.48	0.49	0.49		
Testing	0	0.08	0.07	0.08	0.16	0.97
	1	0.19	0.21	0.20		
	2	0.35	0.36	0.35		
	3	0.38	0.39	0.38		
	4	0.52	0.50	0.51		

Table A2: Per-category Classification Metrics with Overall Cohen's κ and Probability Correlation

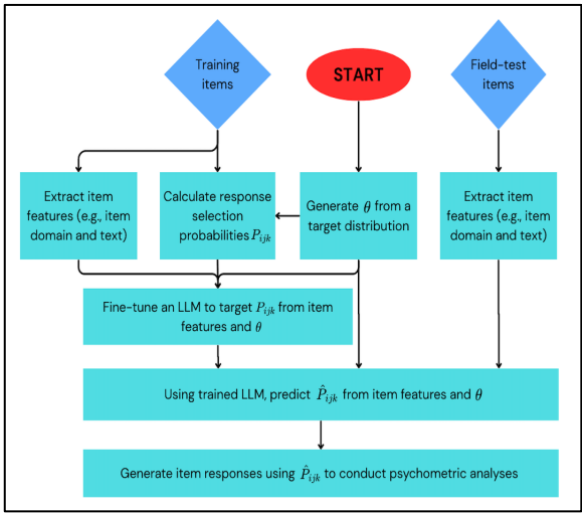

Figure A1: AI-Based Field-Test Data Generation Pipeline

An example of DESSA items:

Domain: Self-Awareness

I can recognize my strengths.

0. Never
1. Rarely
2. Sometimes
3. Often
4. Almost Always

Text consumed by the LLM model after processing item features data:

Domain: Self-Awareness Item: I can recognize my strengths.

Figure A2: DESSA Item Example and Corresponding Preprocessed LLM Input Text