

Prior-Informed 3PL Calibration: Reducing Sample Size via Predictive Modeling

Ruitao Liu
ACT
ruitao.liu@act.org

Aixin Tan
University of Iowa
aixin-tan@uiowa.edu

Abstract

This simulation study evaluates whether predictive item-parameter priors can reduce respondent requirements for 3PL IRT calibration. Across twenty seven prior-quality configurations and seven sample sizes (400–1600), six configurations were able to match a 2000-respondent baseline at 400 respondents, yielding an 80% sample-size reduction when informative priors were properly incorporated.

1 Introduction

Item calibration is a foundational step in the development and maintenance of educational assessments. Under item response theory (IRT), accurate parameter estimation enables reliable test construction, scoring, equating, and adaptive administration (Lord, 1980). For the three-parameter logistic (3PL) model, this task is especially demanding because three item properties must be estimated jointly: discrimination (a), difficulty (b), and a lower asymptote commonly interpreted as a pseudo-guessing parameter (c) (Birnbaum, 1968). The guessing parameter is notoriously difficult to recover even under large sample conditions, because stable estimation depends on sufficient responses from low-ability examinees in the lower tail of the ability distribution (Baker and Kim, 2004).

In practice, calibration quality generally improves as the number of available respondents increases. Standard recommendations suggest a minimum of 500–1000 respondents for the 3PL model under maximum likelihood estimation (Hulin et al., 1982; Stone, 1992), with larger samples preferred for operational use. Yet in many testing programs, assembling large calibration samples is costly and time-consuming.

Bayesian and penalized-likelihood methods offer a natural approach to small-sample item calibration. By incorporating prior information about item parameters, these methods can compensate

for limited response data and improve the stability and accuracy of parameter estimates (Swaminathan and Gifford, 1985, 1986).

The practical challenge is obtaining accurate, item-specific prior information. Recent advances in natural language processing (NLP) and machine learning provide a promising source: item difficulty modeling (IDM), or more generally item parameter modeling, can predict item parameters directly from item text and metadata (AlKhuyaey et al., 2024; Benedetto et al., 2023; Ferrara et al., 2022).

Given an IDM trained on an external item pool, a practical two-step approach is to predict parameters for new items and then use these predictions to construct informative priors for calibration with a smaller number of student responses. Recent studies demonstrate the potential of this approach but also its current limitations. Ulitzsch et al. (2025) found that state-of-the-art IDM predictions yielded little improvement over priors based only on historical item-parameter distributions, while simulations showed that sufficiently improved predictions could substantially reduce calibration sample size. Han and Rijmen (2026) demonstrated empirically that LLM-based predictions can be substantially refined with relatively few student responses. Together, these findings shift the practical question from whether predictive priors can help to *how accurate they must be, for which item parameters, to support a desired reduction in calibration sample size*.

Ulitzsch et al. (2025) explicitly recommend that testing programs evaluate IDM prediction quality through program-specific simulations rather than using their results as look-up tables, because the combinations of prediction accuracy required for sample size reduction are context-dependent. The present study directly answers that call. We compare reduced-sample MAP calibration using IDM-informed priors with a standard 2000-respondent EM calibration. IDM prediction errors for a , b ,

and c are varied separately, allowing us to determine which improvements in IDM accuracy translate into operational reductions in respondent requirements. Performance is evaluated using RMSE, MAE, and Pearson correlation, with a non-inferiority criterion providing a formal rule for determining whether reduced-sample calibration maintains baseline-level performance.

Our results reveal that RMSE and MAE are well controlled even at small sample sizes, whereas correlation between parameter estimations and truth, particularly for c , is the primary obstacle to matching the baseline. Improvements in the c prior have the greatest operational impact, followed by a , while improvements in b contribute relatively little under the conditions studied. With sufficiently improved c and a predictions, calibration with only 400 respondents matches the 2000-respondent baseline under our primary criterion, corresponding to an 80% reduction in respondent requirements. These results provide concrete targets for prioritizing future IDM development.

1.1 Research Questions

This study addresses five research questions:

RQ1. How does sample size affect item-parameter recovery under different prior-quality conditions, and does the effect differ across the a , b , and c parameters?

RQ2. For each prior-quality condition, what is the minimum respondent sample size that matches the 2000-respondent baseline under each matching criterion?

RQ3. Which item parameter prior components (a , b , or c) contribute most to enabling sample-size reduction?

RQ4. When a condition fails to match the baseline, which metric or parameter is the limiting factor?

RQ5. What practical recommendations for operational calibration workflows follow from the simulation results?

2 Related Work

2.1 Bayesian IRT Estimation

Bayesian estimation has a well-established theoretical foundation in IRT. Swaminathan and Gifford (1985, 1986) demonstrated that informative priors can substantially improve estimation of the c parameter, whose likelihood surface is notoriously flat at small sample sizes. Baker and Kim (2004) similarly noted that reliable estimation of c

generally requires larger samples than other item parameters, further motivating informative priors.

For Bayesian point estimation, the posterior mean, or expected a posteriori (EAP), is a natural choice but often requires computationally intensive MCMC. As a computationally simpler alternative that often closely approximates EAP, Mislevy (1986) advocated joint maximum a posteriori (MAP) estimation of item parameters after integrating out ability parameters, an approach also commonly referred to as penalized maximum likelihood (PML).

2.2 Item Parameter Predictions as informative priors

The IDM literature has grown substantially with NLP. Settles et al. (2020) showed that item difficulty can be predicted from linguistic features with meaningful accuracy. Benedetto et al. (2021) applied BERT-based models to predict IRT parameters from item text, achieving difficulty correlations in the range of 0.5–0.7. Lalor et al. (2019) proposed using neural network embeddings as a basis for prior specification.

Han and Rijmen (2026) extended this work to operational K–12 ELA assessments using fine-tuned LLMs. They found that LLMs alone can achieve difficulty correlations approaching 0.7. They also showed that $N \geq 10$ responses can substantially refine LLM predictions for estimating difficulty. For estimating discrimination, gains from respondent data were more modest, requiring $N \geq 30$ responses to show meaningful improvement.

2.3 Benchmarking IDM Prediction Quality

Ulitzsch et al. (2025) provide the most direct methodological precursor to the present study. Using a high-stakes admission test with a 1,742-item pool, they found in their study I that PML estimators are computationally practical and have comparable performance as the computationally more intensive MCMC-based posterior mean for estimating Item Characteristic Curve (ICC) and ability. In their Study II, state-of-the-art IDM predictions ($R_b^2 = .23$; $R_a^2 \approx .02$; $R_c^2 \approx .02$) yielded negligible gains over priors based on historical item-parameter distributions. Their Study III identified counterbalancing patterns: meaningful sample size reduction could be achieved with either high-quality b predictions ($R_b^2 = .80$) or good joint predictions for a and c ($R_a^2 \geq .60$, $R_c^2 \geq .40$), evaluated against a 1000-respondent PML default

baseline.

The present study complements and extends this work in three ways. First, we use a non-inferiority matching framework against a 2000-respondent EM baseline, which is more directly actionable for programs modifying standard calibration practice. Second, we manipulate prediction error directly on the parameter scale rather than defining prior quality through R^2 values. Third, we include Pearson correlation as a matching criterion alongside RMSE and MAE, allowing us to evaluate rank-order recovery in addition to error magnitude.

3 Methods

3.1 Simulation Design Overview

The simulation was designed to answer a practical question: can a smaller-sample prior-informed calibration perform comparably to a 2000-respondent calibration without informative priors? The simulation involved four components: (1) generating item response data under a known 3PL model, (2) constructing informative priors from predicted item parameters, (3) obtain MAP of item parameters via optimization algorithms, and (4) evaluating calibration quality relative to both true parameters and the 2000-respondent baseline.

3.2 Item Response Data Generation

Responses were generated for a fixed 40-item test form, where the true values of a,b,c were fixed and known to the oracle, but not to the analyst. For each replication, person abilities were drawn from $\theta_p \sim N(0, 1)$, and binary responses were generated under the standard 3PL model:

$$P(Y_{pi} = 1 \mid \theta_p) = c_i + (1 - c_i) \times \frac{1}{1 + \exp\{-a_i(\theta_p - b_i)\}}, \quad (1)$$

where a_i , b_i , and c_i are the discrimination, difficulty, and guessing parameters for item i . A master response bank consisting of 15 replications $\times$ 2,000 respondents. Within each replication, calibration samples of $N \in \{400, 600, 800, 1000, 1200, 1400, 1600\}$ were formed by taking the first N respondents.

3.3 Baseline Calibration

The baseline condition used all 2000 respondents and estimated item parameters with the **mirt** package (Chalmers, 2012) using standard EM estimation. These estimates do not use informative item

parameter priors and will serve as the reference benchmark for evaluating the reduced-sample prior-informed calibrations.

3.4 Prior Construction

Recall we are in a situation where a predictive model, such as an IDM, has been trained on an external item pool (not necessarily related to the 40 items above), for which the true item parameters are known, as well as their predicts from the IDM. We next use prediction discrepancies observed on this external item pool to guide the construction of priors. Indeed, for discrimination a with range $(0, \infty)$, we define a normal prior on $\log(a)$, with mean $\mu_{a,i} = \log(\hat{a}_i^{(r)}) + \bar{e}_a$ and standard deviation $\sigma_a = \text{sd}(e_{a,i})$. Here, $e_{a,i} = \log(a_i) - \log(\hat{a}_i^{(r)})$ is the log-scale prediction error, a_i is the true discrimination value for item i , $\hat{a}_i^{(r)}$ is the corresponding predicted value from the predictor indexed by an error reduction factor r (see below), $\bar{e}_a$ and $\text{sd}(e_{a,i})$ are the mean and the standard deviation of the prediction errors across items in the external pool. Note $\bar{e}_a$ serves as a bias correction to the predicted prior mean. For difficulty b , a normal prior is specified with mean $\mu_{b,i} = \hat{b}_i^{(r)} + \bar{e}_b$ and standard deviation $\sigma_b = \text{sd}(e_{b,i})$. For guessing c , a normal prior was similarly specified on $\text{logit}(c)$.

For this simulation study, we manipulate the predictor accuracy, and the implied prior quality, by scaling the prediction error for each parameter using a reduction factor $r \in \{0.00, 0.05, 0.10\}$:

$$\hat{x}_i^{(r)} = x_i + (1 - r) (\hat{x}_i^{\text{pred}} - x_i), \quad (2)$$

where x_i denotes the true item parameter and $\hat{x}_i^{\text{pred}}$ denotes the original predicted value. When $r = 0.00$, no adjustment is made, so the prior is based directly on the original predicted value. When $r > 0$, the original prediction is moved toward the true value by the corresponding proportion of its prediction error. For example, $r = 0.05$ reduces the prediction error by 5%, and $r = 0.10$ reduces the prediction error by 10%. Thus, larger values of r represent more accurate prior information while preserving the general pattern of the original predictions. Crossing three reduction levels for each of a , b , and c produces $3^3 = 27$ prior configurations (labeled 1–27). The magnitude of these manipulations is intentionally modest, reflecting realistic near-term improvement horizons for predictive models in operational programs.

3.5 MAP Estimation

Prior-informed calibration was implemented in R using a custom MAP estimator with the `torch` package, building on our recently developed Simulated Gradient Ascent Method (SGAM; Liu, 2025). SGAM estimates item parameters by maximizing the log posterior through stochastic gradient ascent over Monte Carlo-approximated marginal likelihoods, offering a computationally efficient alternative to EM-based and MCMC-based Bayesian estimators. The estimator maximizes:

$$\log p(\mathbf{Y} \mid \boldsymbol{\eta}) + \log p(\boldsymbol{\eta}), \quad (3)$$

where $\mathbf{Y}$ is the response matrix and $\boldsymbol{\eta}$ is the vector of item parameters including a , b , and c . The marginal log-likelihood was approximated by Monte Carlo integration with 200 ability draws per respondent per iteration from $N(0, 1)$. Optimization used 450 iterations, a learning rate of 0.0025, and full-batch gradient updates. Final estimates were taken from the iteration with the highest recorded log-posterior value. Item parameters were box-constrained: $a \in [0.01, 3]$, $b \in [-4, 4]$, $c \in [0.10, 0.40]$.

3.6 Evaluation Metrics and Matching Criteria

For each replication and condition, three metrics were computed per parameter: root mean squared error (RMSE), mean absolute error (MAE), and Pearson correlation (PC) between estimated and true item parameters across items. RMSE and MAE measure overall error magnitude; Pearson correlation measures rank-order recovery.

A prior-informed condition was considered to *match* the baseline if its performance was non-inferior under a paired replication-level comparison using one-sided confidence intervals ($\alpha = 0.05$) with margins of 5 percent of the baseline value for RMSE and MAE and .05 for Pearson correlation. For RMSE and MAE, lower values were better; for Pearson correlation, higher values were better. Two matching criteria were defined:

- **Criterion 1 (C1):** non-inferior RMSE and MAE for all three parameters (a , b , c).
- **Criterion 2 (C2):** non-inferior RMSE, MAE, and Pearson correlation for all three parameters.

C2 was treated as the primary operational criterion because it requires both controlled error mag-

nitude and preserved item rank ordering—essential for test assembly, item banking, and equating.

4 Results

4.1 Baseline Performance

The 2000-respondent EM calibration established the following benchmark across 15 replications:

a : RMSE = 0.221, MAE = 0.170, PC = 0.727;

b : RMSE = 0.628, MAE = 0.453, PC = 0.776;

c : RMSE = 0.131, MAE = 0.107, PC = 0.319.

Note that the Pearson correlation for c was substantially lower than that for a and b , consistent with the well-documented estimation difficulty for the guessing parameter even with all 2000 respondents.

Table 1 places the baseline alongside no-reduction prior-informed conditions ($r = 0$ for all parameters) at sample sizes 400–1000. MAP estimation with informative priors consistently produced *lower* RMSE and MAE than the EM baseline across all sample sizes, consistent with Ulitzsch et al.’s (2025) observation that even modestly informative priors improve error-magnitude metrics. However, Pearson correlations for c remained substantially below the baseline, indicating that MAP controls average error but struggles to preserve the rank ordering of c values when the prior is insufficiently accurate.

4.2 RQ1: Effect of Sample Size on Parameter Recovery

Increasing sample size improved parameter recovery across all prior configurations, but the nature of that improvement varied by parameter. For a and b , RMSE and MAE already met the baseline-matching criterion at 400 respondents across all prior-quality conditions, and improvements with additional respondents were modest. For c , RMSE and MAE were already low at small sample sizes, consistent with the regularizing effect of the informative prior. However, Pearson correlation for c was substantially below the baseline at 400 respondents and improved only slowly. This dissociation between error magnitude and rank-order recovery is a key finding: a calibration can appear accurate on average deviation metrics while still failing to order items correctly.

4.3 RQ2: Minimum Sample Sizes under Each Matching Criterion

Under C1, all 27 prior configurations matched the baseline on RMSE and MAE at 400 respondents,

Table 1: Calibration performance of the baseline estimator (via EM) using all $N = 2,000$ responses versus IDM-informed estimators using samples of varying sizes, N . The IDM predictions here have the best achievable accuracy in this context (prediction reduction factor $r = 0$). Lower RMSE and MAE and higher PC indicate better performance.

Estimator	N	a			b			c		
		RMSE	MAE	PC	RMSE	MAE	PC	RMSE	MAE	PC
Baseline	2000	.221	.170	.727	.628	.453	.776	.131	.107	.319
	400	.196	.152	.674	.445	.343	.868	.082	.066	.205
MAP based on	600	.191	.147	.701	.440	.336	.879	.081	.065	.213
IDM-informed prior	800	.185	.143	.716	.429	.330	.885	.081	.065	.217
	1000	.180	.139	.733	.425	.324	.892	.081	.065	.244

Table 2: Minimum sample size for representative prior configurations under both matching criteria. “No match” = failed to match at any $N \leq 1600$.

Configuration	r_a	r_b	r_c	$N_{\min}$ C1	$N_{\min}$ C2
1	0.00	0.00	0.00	400	No match
9	0.10	0.10	0.00	400	No match
10	0.00	0.00	0.05	400	600
11	0.05	0.00	0.05	400	400
14	0.05	0.05	0.05	400	400
16	0.00	0.10	0.05	400	600
19	0.00	0.00	0.10	400	600
20	0.05	0.00	0.10	400	400
23	0.05	0.05	0.10	400	400
27	0.10	0.10	0.10	400	400

indicating that average error was well controlled across the tested prior-quality conditions. However, C1 provides little operational discrimination because every prior configuration passes at the smallest tested sample.

Under C2, results were far more informative. The number of matching prior configurations at each sample size was: 6 of 27 at 400; 9 of 27 at 600; 9 of 27 through 1200; 15 of 27 at 1400; and 15 of 27 at 1600. Eleven prior configurations failed to match at any tested sample size. The plateau from 600 to 1200 is particularly notable: conditions that cannot match at 600 also cannot match at 800, 1000, or 1200. In this region, increasing sample size yields no benefit, the possible path to improvement is improving prior quality.

Table 2 summarizes minimum matching sample sizes for a representative set of prior configurations. Prior configuration 1 (no improvement on any parameter) never matches the baseline, while prior configuration 27 (10% error reduction on all parameters) matches at 400 respondents, illustrating the importance of prior quality for reducing calibration sample size.

4.4 RQ3: Relative Importance of Prior Components

A clear hierarchy of prior importance emerged. The c prior was decisive: when $c_{\text{reduction}} = 0.00$, none of the nine prior configurations in that group matched the baseline at any tested sample size, regardless of a or b prior strength. This holds even for prior configuration 9, which combines maximum a and b improvement with no c improvement. This finding directly is consistent with Ullrich et al.’s (2025) observation that current IDM predictions were particularly weak for c ($R_c^2 \approx .02$), suggesting that limited c -prediction accuracy may have contributed to their limited calibration gains.

When $c_{\text{reduction}} = 0.10$, conditions with $a_{\text{reduction}} \geq 0.05$ matched at 400 respondents; conditions with $a_{\text{reduction}} = 0.00$ required 600. The b prior had the smallest operational effect—changing $b_{\text{reduction}}$ from 0.00 to 0.10 rarely changed the minimum matching sample size. The practical ranking of prior importance was: $c \gg a > b$.

4.5 RQ4: Diagnosing Matching Failures

Across all 189 prior-informed conditions, there were zero failures on RMSE or MAE for any parameter, and zero failures on b Pearson correlation. Every failure under C2 was attributable to one of only two metrics: a Pearson correlation (9 failures, concentrated at 400 respondents) and c Pearson correlation (114 failures, the dominant failure mode at all sample sizes). This confirms that rank-order recovery of item parameters, not error magnitude, is the binding constraint in this framework. From 600 respondents onward, a Pearson correlation no longer caused matching failures; the only remaining obstacle was c Pearson correlation.

4.6 RQ5: Practical Recommendations

Recommendation 1: Adopt C2 as the primary evaluation standard. C1 is satisfied by all conditions and provides no useful discrimination. C2 is the only criterion that identifies when a program’s prediction system is ready to support reduced-sample calibration.

Recommendation 2: Prioritize improvement of the c prior. Without improvement in c prior accuracy, no configuration matched the baseline even at 1600 respondents. Programs should evaluate their predictive models’ ability to rank items by guessing parameters, not just assess difficulty accuracy (Benedetto et al., 2021; Han and Rijmen, 2026; Settles et al., 2020).

Recommendation 3: Target 400 respondents when both c and a priors are strong. When $c_{\text{reduction}} \geq 0.10$ and $a_{\text{reduction}} \geq 0.05$, calibration at 400 respondents matches the 2000-respondent baseline under C2—an 80% reduction.

Recommendation 4: Use 600 respondents as a safer intermediate threshold. When $c_{\text{reduction}} \geq 0.10$ but $a_{\text{reduction}} = 0$, 600 respondents was sufficient for a 70% reduction.

5 Discussion

This study provides simulation-based evidence that predictive item-parameter priors can meaningfully reduce respondent samples for operational 3PL calibration—but only under well-defined conditions. Three main contributions emerge relative to prior work.

Separating error magnitude from rank-order recovery. The consistent finding that RMSE and MAE were easily matched while Pearson correlation was often the binding constraint highlights a limitation of relying solely on error-magnitude metrics. Prior work, including Ulitzsch et al. (2025), has primarily used RMSE-based evaluation; our results suggests that rank-order metrics can provide additional information when evaluating IRT calibration quality in IDM contexts.

The c prior as a decisive operational lever. The strong dependence on c prior quality is consistent with Ulitzsch et al. (2025), who reported very limited predictive accuracy for c ($R_c^2 \approx .02$) and correspondingly limited gains from state-of-the-art IDM predictions. This suggests a more optimistic interpretation of their findings: the framework is sound; current IDM prediction quality for c is simply not yet at the level needed to unlock

sample-size reductions. As prediction accuracy for c improves—whether through fine-tuned LLMs as demonstrated by Han and Rijmen (2026) or other approaches—the sample size reduction identified here may become increasingly achievable.

A non-inferiority framework for calibration decisions. By framing reduced-sample calibration as a baseline-matching problem, this study provides a practical decision framework for evaluating reduced-sample calibration. Ulitzsch et al. (2025) recommend that programs run their own simulations mirroring local conditions; the present non-inferiority framework offers a concrete statistical methodology for doing so.

Several limitations should be noted. The simulation was conducted on a single 40-item test form; item pool characteristics such as the range of true c values, item discrimination, and test length may influence results. The prior quality manipulations were modest (maximum 10% error reduction), chosen to reflect realistic near-term improvements. Finally, the non-inferiority margins were fixed; sensitivity analyses varying these margins would strengthen the conclusions.

6 Conclusion

This study shows that IDS-informed priors for item parameters could support substantial reductions in 3PL calibration sample size, down to 400 respondents from 2000, if the IDM prediction accuracy for the guessing parameter, c , meets certain threshold. Indeed, c emerged as the primary bottleneck: its rank-order recovery was the dominant source of matching failures across conditions, and no configuration succeeded without adequate c prior quality. These findings help identify improving c prediction as an important requirement for achieving sample-size reductions. The non-inferiority matching framework offers a practical approach for evaluating whether a prediction system can support reduced-sample calibration.

Limitations

This simulation was conducted on a single test form with 40 items and modest prior quality manipulations (up to 10% error reduction). Results may differ for forms with different item pool characteristics. The non-inferiority margins used in the matching comparisons were fixed; sensitivity analyses are warranted.

References

- Samah AlKhazaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Frank B. Baker and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*, 2nd edition. Marcel Dekker.
- Luca Benedetto, Paolo Cremonesi, and Michele Andreoletti. 2021. Application of a BERT-based scoring function for predicting item difficulty in medical licensing examinations. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*.
- Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Alessio Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.
- Allan Birnbaum. 1968. Some latent trait models and their use in inferring an examinee’s ability. In Fredric M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Addison-Wesley.
- R. Philip Chalmers. 2012. [mirt: A multidimensional item response theory package for the R environment](#). *Journal of Statistical Software*, 48(6):1–29.
- Steve Ferrara, Jeffrey T. Steedle, and Robert S. Frantz. 2022. Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, 35(3):237–253.
- Suhwa Han and Frank Rijmen. 2026. A sequential framework for item parameter modeling: Bayesian refinement of LLM-based priors. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Denver, CO.
- Charles L. Hulin, Robin I. Lissak, and Fritz Drasgow. 1982. Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6(3):249–260.
- John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4240–4250.
- Ruitao Liu. 2025. Rethinking IRT calibration: Simulated gradient ascent method (SGAM). Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Denver, CO.
- Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.
- Robert J. Mislevy. 1986. Bayes modal estimation in item response models. *Psychometrika*, 51(2):177–195.
- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.
- Clement A. Stone. 1992. Recovery of marginal maximum likelihood estimates in the two-parameter logistic model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16(1):1–16.
- Hariharan Swaminathan and Janice A. Gifford. 1985. Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50(3):349–364.
- Hariharan Swaminathan and Janice A. Gifford. 1986. Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51(4):589–601.
- 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*.