

Evaluating Neural Network-Based IRT Estimation in Multistage Adaptive Testing with Small Item Banks

Seong Eun Hong

Center for Applied Linguistics
shong@cal.org

Abstract

Neural network (NN)-based item response theory (IRT) estimation offers a potential alternative for calibrating items when operational samples are small, but its accuracy may depend on how well simulated training data represents later applications. This study investigates the performance of NN-based IRT estimation in multistage adaptive language assessment with small item banks and short tests. Results showed that accuracy improved with larger samples and more training information. Moderate distribution shifts were largely accommodated under higher iteration conditions, whereas larger shifts continued to reduce estimation accuracy.

1 Introduction

Accurate item parameter estimation is essential for item response theory (IRT), but calibration becomes challenging when only small samples are available. Limited response information can result in unstable parameter estimates, and this becomes more challenging in adaptive testing, where responses are sparse because examinees encounter different subsets of items (Ban et al., 2002). In multistage testing (MST), routing further structures item exposure by assigning examinees to modules targeted to different proficiency levels. Consequently, interactions between examinee groups and modules can create additional challenges for item calibration (Cai et al., 2021). Recent work using an operational MST program similarly showed that calibrating new items with small samples remains an important practical

problem and that calibration performance depends on the estimation and data-collection design employed (Guo et al., 2024).

Machine-learning approaches offer an alternative way to address limitations of conventional estimation. Deep learning has increasingly been incorporated into psychometric modeling, including computationally efficient estimation for complex item factor models (Urban & Bauer, 2021). More recently, Belov et al. (2026) introduced a supervised neural network (NN) approach for estimating IRT parameters in small samples. Instead of estimating parameters directly from a likelihood function, the method derives features from simulated response data and trains an NN to map these features to known item parameters. Their simulation results showed that NN-based estimators could produce accurate parameter estimates with small samples and outperform Bayesian MCMC estimation under the conditions examined. Related work has also explored incorporating machine-learning-based item parameter predictions into small-sample calibration, emphasizing the potential value of prediction when calibration information is limited (Ulitzsch et al., 2026).

A potential limitation of supervised NN estimation, however, is its dependence on the conditions represented in the training data. Belov et al. (2026) identified the characteristics and distributions of simulated training data as an important area for further investigation. Although training and test distribution mismatch is not unique to adaptive testing, it may be relevant in MST because adaptive routing intentionally exposes different examinee groups to item sets

with different difficulty characteristics. An NN trained on one difficulty distribution may therefore be applied to operational data whose difficulty distribution differs because of modules, item-pool composition, or administration conditions. Evaluating this type of distributional shift provides evidence about whether NN-based parameter estimation generalizes beyond the conditions on which the model was trained.

2 Study Purpose

This study evaluates NN-based item difficulty estimation under MST conditions using relatively small Listening and Reading item banks. The simulation varies calibration sample size, the amount of simulated training information, and the degree of mismatch between the item difficulty distributions represented in the training and test data. This study addresses the following research questions:

RQ1. How do calibration sample size, item pool size, and the amount of simulated training information affect the accuracy of NN-based item difficulty estimates under MST conditions?

RQ2. To what extent does a mismatch between the item difficulty distributions represented in the training and test data affect the accuracy of NN-based Rasch item difficulty estimates under MST conditions?

3 Method

3.1 Data

The simulation design was based on operational characteristics of the WIDA ACCESS Listening and Reading assessments. Both domains use an MST design and are administered across five grade clusters: G01, G23, G45, G6-8, and G9-12. The Listening domain consists of six or eight stages, whereas Reading consists of eight or ten stages. The first two stages serve as routing stages. Based on performance during routing, examinees are assigned to Tier A, B, or C. Examinees assigned to Tiers B and C may progress to later stages that are not administered to students routed to a lower tier.

The number of items administered therefore differs according to the student's route. Listening tests contain either 18 or 24 administered items, and Reading tests contain either 24 or 30 items. Item difficulty parameters for both domains were

estimated using the Rasch model and vertically scaled across grade clusters. These operational item parameter estimates served as the basis for constructing the simulation item pools. Table 1 presents the descriptive statistics of item difficulty parameters for Listening and Reading domain.

Grade	Item	Mean	P25	P75	SD
Listening					
G01	54	-1.69	-2.40	-0.59	1.23
G23	54	-0.37	-1.78	0.99	1.71
G45	54	0.22	-0.70	1.06	1.12
G68	54	0.30	-0.55	1.01	1.10
G91	54	1.15	0.34	1.94	1.12
Reading					
G01	72	-1.40	-2.16	-0.37	1.31
G23	72	-0.62	-1.35	0.29	1.16
G45	72	0.11	-0.80	1.11	1.23
G68	72	1.15	0.14	2.24	1.34
G91	72	1.61	0.96	2.47	1.18

Table 1: Descriptive statistics of item difficulty parameters for Listening and Reading domains across grade clusters

3.2 Simulation Design

The simulation evaluated NN-based estimation of item difficulty parameters under combinations of four factors: item pool size, calibration sample size, amount of simulated training information, and degree of training-test distribution mismatch. The factor levels followed the operational item-bank characteristics and extended the simulation framework used in previous NN-based IRT research.

The item pool factor contained two levels corresponding to the two operational domains: $P = 270$ for Listening and $P = 360$ for Reading. Calibration sample size was varied across seven levels, $S = 10, 50, 100, 150, 200, 250$, and 300 . This range was selected to represent extremely limited through moderate calibration samples. The amount of simulated training information was manipulated through the number of iterations used to construct the training data: $I = 5, 10, 15$, and 20 . In this study, an iteration represents an additional simulated MST response dataset generated from the item pool and therefore increases the amount of response-pattern information available for training the NN.

The final factor represented the correspondence between the item difficulty distributions used for

NN training and evaluation. Three conditions were examined. In the matched condition, test item difficulties were generated from the same difficulty distribution represented in the training data. In the moderate-shift condition, the mean test-item difficulty was shifted by 0.5 logits while in the extreme-shift condition, the mean test-item difficulty was shifted by 1.0 logit.

3.3 Training and Test Data Generation

For each simulation condition, training and test datasets were generated using the same MST routing structure. In Listening, six routing items were administered first, followed by 12 additional items for Tier A or 18 additional items for Tiers B and C. Reading data followed the corresponding operational MST map and domain-specific test lengths. Items were sampled from the relevant operational pool so that simulated examinees encountered only the items associated with their assigned route rather than the complete bank.

For each simulated examinee, latent proficiency was drawn from a standard normal distribution, $\theta \sim N(0, 1)$. Binary responses were then generated under the Rasch model as a function of examinee proficiency and item difficulty. Response-derived variables and normalized scores used as NN inputs were constructed following the supervised-learning procedure described by Belov et al. (2025). Repeating this process according to the assigned iteration condition produced the training information used to fit the estimator.

Independent test datasets were generated with the same response model, proficiency distribution, routing structure, and test lengths. The only systematic difference was the item-difficulty distribution: it was unchanged in the matched condition and shifted upward by 0.5 or 1.0 logit in the moderate- and large-shift conditions, respectively. This design holds the remaining data-generating features constant. One hundred Monte Carlo replications were attempted for each condition.

3.4 Neural Network Estimation

The NN-based estimators were implemented in Python using the TensorFlow library (TensorFlow Developers, 2023) for model training and item parameter estimation. Hyperparameter settings followed those used in Belov et al. (2025). For training, the generating item difficulties were used as item labels, and each item's response row was

concatenated with the sorted normalized raw scores of the examinee sample to form a feature vector. The trained NN was then applied to feature vectors from independent test datasets to obtain item difficulty estimates.

The primary distinction from conventional IRT calibration is that the trained network does not re-estimate a likelihood for each test dataset. Instead, it applies a previously learned mapping from response-derived characteristics to item difficulty. This makes training data coverage a substantive part of the estimation problem: if the response patterns associated with new items fall outside the range represented during training, prediction error may increase even when the underlying Rasch model is correctly specified.

Estimation accuracy was evaluated using root mean squared error (RMSE) between the generating item difficulties and the NN estimates. RMSE summarizes the magnitude of parameter-recovery error on the Rasch difficulty scale, with smaller values indicating more accurate recovery. For each condition, RMSE was computed across evaluated items and summarized across Monte Carlo replications.

4 Results

Figures 1 and 2 compare the matched, moderate-shift, and large-shift conditions. For clarity, only lowest and highest iteration conditions ($I = 5$ and $I = 20$), are displayed. As expected, the large-shift condition generally produced highest RMSE, particularly when the number of training iterations was low. In contrast, the moderate-shift condition with $I = 20$ followed the matched-condition results closely across most sample sizes in both Listening and Reading. This pattern suggests that additional training information can substantially reduce the effect of a moderate mismatch between the training and test item-difficulty distributions. However, the same degree of recovery was not evident under the large-shift condition, where RMSE remained higher even with $I = 20$.

Overall, the results indicate that NN-based item parameter estimation benefits from both larger calibration samples and greater amount of simulated training information. More importantly, the effect of distributional mismatch appears to depend on its magnitude: moderate shifts can be largely accommodated when sufficient training information is available, whereas larger shifts continue to reduce estimation accuracy.

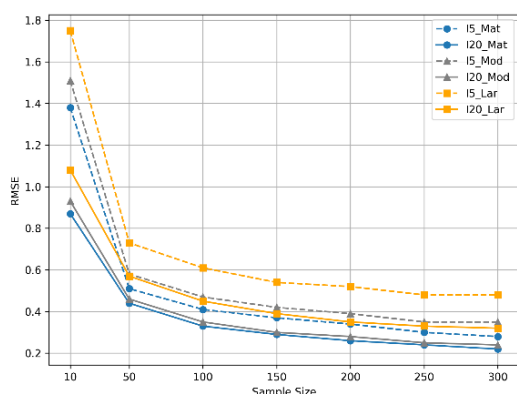


Figure 1: RMSE of item difficulty estimates for Listening across matched, moderate-shift, and large-shift conditions.

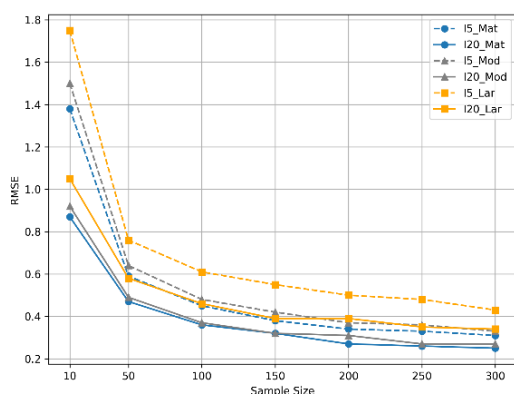


Figure 2: RMSE of item difficulty estimates for Reading across matched, moderate-shift, and large-shift conditions.

5 Conclusion

This study evaluated the performance and robustness of NN-based Rasch item difficulty estimation in an MST setting with relatively small operational item banks. Estimation accuracy improved as sample size and the amount of simulated training information increased, with the greatest gains occurring at smaller sample sizes and diminishing improvements at larger samples.

The findings also clarify the role of training–test distribution mismatch. Although estimation accuracy generally declined as the test-item difficulty distribution moved away from the training distribution, this effect was not uniform. Under the moderate-shift condition, estimates based on the higher number of training iterations

closely approximated those obtained under the matched condition, suggesting that sufficient training information can provide meaningful robustness to moderate distributional changes. In contrast, a larger distribution shift continued to produce noticeably higher RMSE, even under the high iteration condition.

These findings suggest that NN-based IRT estimation may be practical for MST assessments with limited calibration samples, particularly when simulated training data are sufficiently rich and reasonably representative of operational item difficulty. The results also indicate that exact correspondence between training and operational distributions may not always be necessary; moderate differences may be tolerable when the NN is trained with adequate information. However, substantial distributional shifts remain an important limitation and should be considered when designing training simulations for operational use.

References

- Ban, J. C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, 39(3), 207–218. <https://doi.org/10.1111/j.1745-3984.2002.tb01174.x>
- Belov, D. I., Lüdtke, O., & Ulitzsch, E. (2026). A supervised learning approach to estimating IRT models in small samples. *British Journal of Mathematical and Statistical Psychology*, 79(1), 66–94. <https://doi.org/10.1111/bmsp.12396>
- Cai, L., Albano, A. D., & Roussos, L. A. (2021). An investigation of item calibration methods in multistage testing. *Measurement: Interdisciplinary Research and Perspectives*, 19(3), 163–178. <https://doi.org/10.1080/15366367.2021.1878778>
- Guo, H., Johnson, M. S., McCaffrey, D. F., & Gu, L. (2024). Practical considerations in item calibration with small samples under multistage test design: A case study. *ETS Research Report Series*, 2024(1), 1–21. <https://doi.org/10.1002/ets2.12376>
- TensorFlow Developers. (2023). *TensorFlow* (v2.21.0). Zenodo. <https://zenodo.org/records/18894642>
- Ulitzsch, E., Belov, D., Lüdtke, O., & Robitzsch, A. (2026). Using item parameter predictions for reducing calibration sample requirements—A case study based on a high-stakes admission test. *Journal*

of *Educational Measurement*, 63(1).
<https://doi.org/10.1111/jedm.12426>

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, 86(1), 1–29.
<https://doi.org/10.1007/s11336-021-09748-3>