Rasch Model Sample Size Requirements Under an IRT Pre-Equating Design

Kylie N. Gorney[1] and Andrew A. Mroch[2]

[1]Department of Educational Psychology, University of Wisconsin-Madison
[2]National Conference of Bar Examiners

Introduction

- In an **embedded pretesting design**, pretest items are incorporated into the operational test such that the examinees do not know which items are scored and which are unscored (Schmeiser & Welch, 2006).
- In **item response theory (IRT) pre-equating**, item parameter estimates obtained from previous administrations are used to score examinees in the current administration (Kolen & Brennan, 2014). In particular, these estimates may have been obtained back when the items were pretested. Thus, it is crucial that the pretest item parameter estimates are accurate, else future ability parameter estimates and test scores may be biased.
- Much of the research conducted on IRT pre-equating with an embedded pretesting design has focused on the 3PL model (e.g., Bejar & Wingersky, 1982). However, the **Rasch model** requires fewer examinees for adequate calibration, thereby allowing a higher volume of items to be pretested at once. As a result, administrators are given more options to choose from when selecting which items will eventually become scored.
- **The purpose of this study was to determine the minimum sample size needed to produce accurate item parameter estimates, and in turn, accurate ability parameter estimates, under an IRT pre-equating design using the Rasch model.**

Method

Real Data

- Five test forms, each comprised of 50 common (scored) items, 10 unique (pretest) items, and administered to 2,000 examinees
- Fit the Rasch model to the data using Winsteps (Linacre, 2017), which produced baseline item parameter estimates and baseline ability parameter estimates

Simulated Data

- Simulated 2,000 examinees per test form using the baseline parameter estimates

Item Parameter Estimation

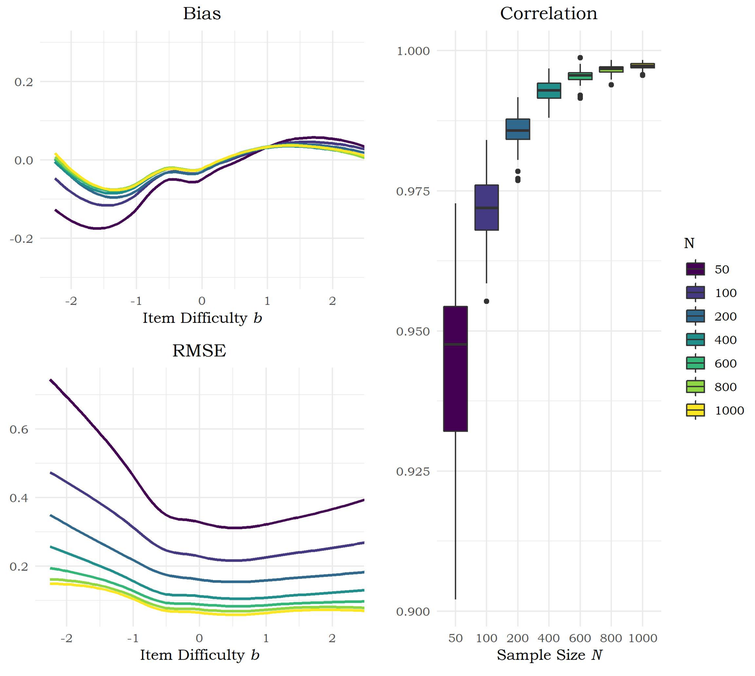
- Sample sizes: $N = 50, 100, 200, 400, 600, 800, 1000$ examinees per test form
- Fixed the item parameter estimates of the 50 scored items to their baseline values, then concurrently calibrated the 50 pretest items using the sample data
- Compared pretest item parameter estimates to their baseline values

Ability Parameter Estimation

- Simulated a hypothetical, future test using the baseline pretest item parameter estimates (mean = 0.12, SD = 1.00) and the baseline ability parameter estimates (mean = 0.76, SD = 0.66)
- Fixed the pretest item parameter estimates to their sample values, then estimated the ability parameters
- Compared ability parameter estimates to their baseline values

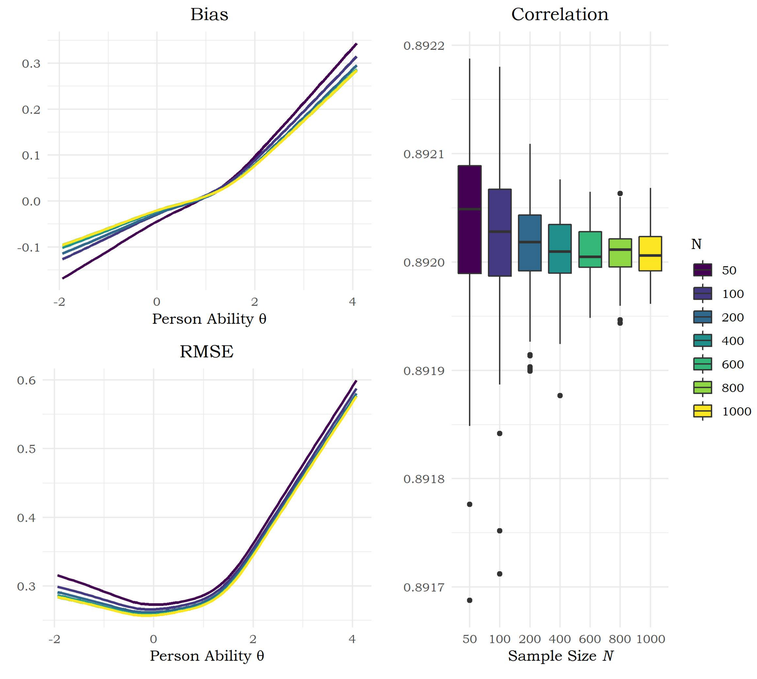
Item Parameter Recovery

Since the real and simulated data yielded similar results, only the results of the simulated data are presented here. The plots below show the average bias and root mean squared error (RMSE) across 100 replications, and the average correlations between the pretest item parameter estimates and their baseline values.



Ability Parameter Recovery

The pretest item parameter estimates were used in a hypothetical, future test to estimate the ability parameters. The accuracy of these estimates is examined here.



References

- Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, *6*(3), 309–325.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7
- Linacre, J. M. (2017). *Winsteps* ® (Version 4.0.1) [Computer software]. https://www.winsteps.com
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). American Council on Education and Praeger Publishers.

Conclusion

Item Difficulty & Person Ability

- As expected, the bias and RMSE increased at the extreme ends of the scales. In other words, the parameters for items of medium difficulty and persons of medium ability were estimated with the highest accuracy.

Sample Size

- Larger sample sizes yielded more accurate pretest item parameter estimates, on average. For this data set in particular, setting the minimum sample size at 400 examinees appears to be sufficient for adequate item parameter recovery.
- The ability parameter estimates were far less affected by sample size. Even using a sample of 50 examinees to estimate pretest item parameters produced fairly accurate ability parameter estimates, in turn.