**Paper SD-10**

# Evaluating the Performance of the SAS® GLIMMIX Procedure for the Dichotomous Rasch model: A Simulation Study

Isaac Li, University of South Florida, Tampa, FL
Yi-Hsin Chen, University of South Florida, Tampa, FL
Jeffrey D. Kromrey, University of South Florida, Tampa, FL

## ABSTRACT

A simulation study was devised to evaluate the accuracy and precision of the GLIMMIX procedure when fitting the dichotomous Rasch model. The evaluation reviewed technicalities including item parameter recovery, standard error estimates, unstandardized and standardized fit indices produced by GLIMMIX. Factors manipulated in this study were test length (10, 20, 40, 60, and 80) and sample size (100, 200, 400, 800, 1000, 1500, and 2000). The generating item difficulty parameters were symmetrically and evenly distributed between -3 and 3 in logit scale so that the mean item difficulty of the test is 0. The person ability parameters were generated from a normal distribution with a mean of 0 and standard deviation of 1. The following statistics were applied to evaluate the performance of the GLIMMIX procedure: bias, sampling variance of the estimates, average error variance, and descriptive statistics (mean, variance, minimum, and maximum) for INFIT and OUTFIT and standardized INFIT and OUTFIT. The results indicated that SAS GLIMMIX procedure for the dichotomous Rasch model provided biased estimates for smaller sample size and shorter tests. To utilize this analytical tool, applying it to tests longer than 20 items and samples greater than 200 persons is recommended.

Keywords: RASCH, SIMULATION, BASE SAS, SAS/STAT, GLIMMIX

## INTRODUCTION

The Rasch model and its extended models have been widely applied in many fields of research. Practitioners often resort to specialized computer program such as BIGSTEPS and WINSTEPS (BIGSTEPS is the MS-DOS version of WINSTEPS), ConQuest, Facets, LPCM, Quest, RASCAL, RUMM2020, RSP, T-Rasch, and WINMIRA. There is a need for generalized statistical software to be able to perform this kid of complex modeling.

Originally from a macro, the GLIMMIX procedure is a new procedure in SAS/STAT software. It was an add-on for the SAS/STAT product in SAS 9.1 on the Windows platform and has seen improvements in its recent SAS 9.3 version. PROC GLIMMIX performs estimation and statistical inference for generalized linear mixed models (GLMMs). A generalized linear mixed model is a statistical model that extends the class of generalized linear models (GLMs) by incorporating normally distributed random effects. A GLM can be defined in terms of a response distribution for a dependent variable from the exponential family of distributions of several model components.

The GLIMMIX procedure fits generalized linear mixed models based on linearizations. The default estimation method in PROC GLIMMIX for models containing random effects is a technique known as restricted pseudo-likelihood (RPL) estimation. PROC GLIMMIX extends the SAS mixed model tools in a number of ways, including fitting models to multivariate data in which observations do not all have the same distribution or link. The focus of this study is to fit a standard dichotomous Rasch model with GLIMMIX and evaluate its capabilities in terms of item parameter recovery, standard error estimates, and fit statistics.

## ANALYSIS

Parameter recovery analysis looks at whether GLIMMIX can recover the generating parameters accurately. If the empirical mean of the estimates across replications is different from the generating value in a way that is statistically significant, the estimator is said to be biased. Standard error of the estimates is also of concern as it reflects the variability of the estimates across replications. Fit statistics are developed to screen misfitting items or persons. If fit statistics are incorrect, a misfitting item (person) may not be located correctly, or an appropriate item (person) may be identified incorrectly as a misfitting item (person). Multiple fit statistics exist for the Rasch model but this study focuses on the INFIT and OUTFIT mean squares and their *t* transformed counterparts.

To assess the estimation bias, the difference between the mean estimate across all replications and the generating value was computed as

$$Bias(\hat{\zeta}) = \left( \sum_{k=1}^{N_{rep}} \hat{\zeta}_k / N_{rep} \right) - \zeta$$

where $\zeta$ denotes the generating value, $\hat{\zeta}_k$ denotes its estimate in the $k^{th}$ replication, and $N_{rep}$ the number of replications used in the simulation. The sampling variance of the estimates across all replications was computed as

$$SV(\hat{\zeta}) = \sum_{k=1}^{N_{rep}} \left( \hat{\zeta}_k - \hat{\zeta}_r \right)^2 / (N_{rep} - 1)$$

where $\hat{\zeta}_r$ denotes the mean of the estimates over replications. To test whether the estimator $\hat{\zeta}_r$ was biased, we computed

$$Z(\hat{\zeta}) = \frac{Bias(\hat{\zeta})}{\sqrt{SV(\hat{\zeta})/N_{rep}}}$$

The zeta statistic may be referred to the standard normal distribution to test its statistical significance. If the null hypothesis was rejected, the estimator $\hat{\zeta}_r$ was declared to be a biased estimator of $\zeta$ in the sense of statistical significance.

The standard error estimates were squared to form the error variance estimates. The error variance estimates were then averaged across all replications to form the average error variance estimate

$$AEV(\hat{\zeta}) = \sum_{k=1}^{N_{rep}} SE(\hat{\zeta}_k)^2 / N_{rep}$$

where $SE(\hat{\zeta}_k)$ is the standard error estimate of parameter $\zeta$ in the $k^{th}$ replication. If the standard error estimates were accurate, the ratio of the average error variance estimate over the sampling variance would approach one. If the ratio was significantly different than one, the standard error might be overestimated.

## METHOD

Simulation was designed to manipulate two independent variables: test length and sample size. Under the dichotomous Rasch model, the test lengths were set to 10, 20, 40, 60, and 80 items. For each test length, the test was created by setting the difficulty parameter of the most difficult item as 3 in logit scale and that of the easiest item as -3. The difficulty parameters of the remaining items were evenly spaced within this range so that the mean item difficulty in logit of the test was always 0.

All examinee samples were generated from *N*(0, 1) using the SAS RAND function. Sizes of these samples included 100, 200, 400, 800, 1,000, 1,500 and 2,000 persons. A SAS program was written to simulate item responses. The data simulating procedure consisted of the following steps. (1) The ability parameters (thetas) for the entire sample were generated and saved. (2) The theta value for person *v* was read into SAS, and along with the pre-defined difficulty parameters of the items within a test they were used to compute the probability of answering the item correctly as well as the cumulative probabilities under the dichotomous Rasch model. (3) These cumulative probability values were compared with a random number from a uniform [0, 1] distribution, generated from the intrinsic SAS random number function. The simulated item response became 1 if the random number was less than or equal to the associated cumulative probability and 0 otherwise.

Therefore there were 35 simulations from five test lengths times seven sample sizes, for each of which 500 replications were made to create 17,500 response data sets. The simulated data sets were analyzed using the GLIMMIX_RASCH macro (Chen et al., 2013), which produces variance and residual estimates in logit scale. Using SAS SQL procedure, these values were then used to calculate person and item INFIT and OUTFIT mean square statistics.

## SIMULATION RESULTS

*Bias and absolute bias*. Bias was calculated as the difference between the estimated values and the true item difficulty parameters used to generate the response data. Smaller biases indicate better item parameter recovery. Figure 1 depicts the scatter plots of calculated biases against true item difficulty parameters which were grouped by

test length and sample size respectively. It is clear that for a test with only 10 or 20 items, estimation bias at the more difficult and easier items was substantial. As test length increases beyond 20 items, such biases appeared to stabilize between -.2 and .2. It is also clear that the significant bias values associated with the 10- and 20-item tests persisted as the sample size went up.

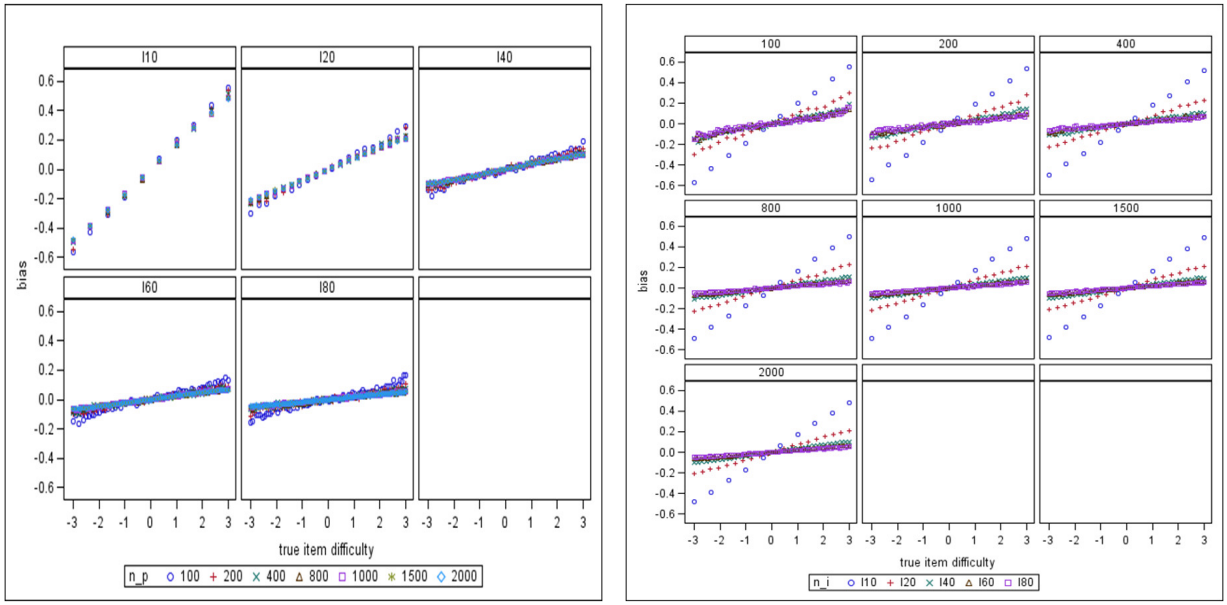


*Figure 1*. Scatter plots of estimation bias against true item parameters

The magnitude of the estimation biases can be looked at by taking their absolute values. Figure 2 plots the average moments of these absolute values (maximum, minimum, mean, and standard deviation) over each simulation condition (test length/sample size). The 20-item test showed the highest bias values and for tests with 40 or more items, the bias estimates became much lower, especially with sample sizes greater than 200. Also, within each test length, there was not much variation with changes in sample size.

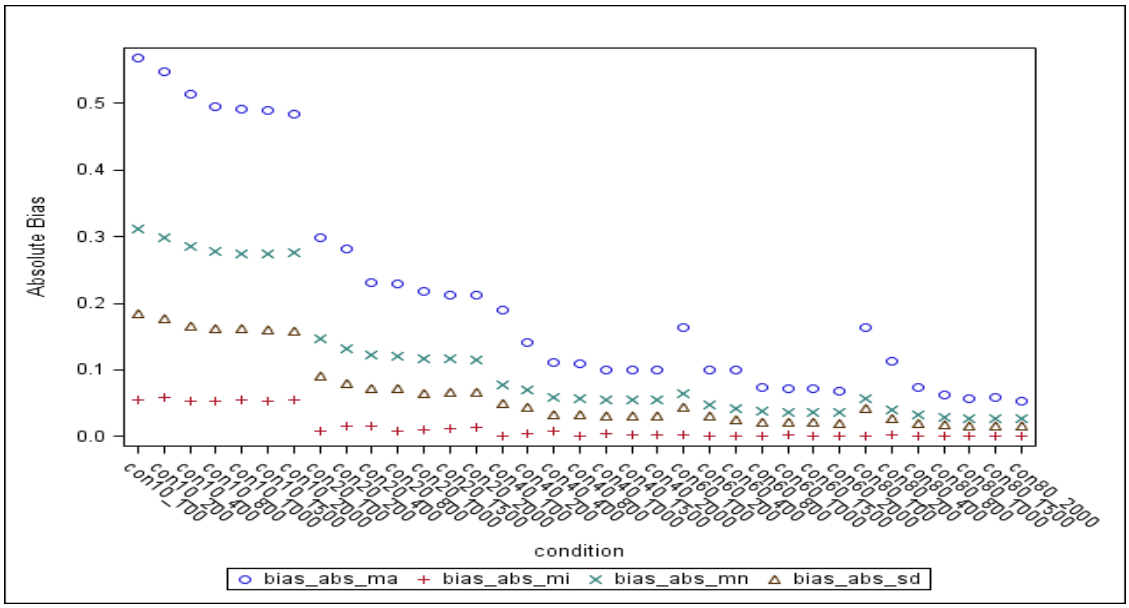


*Figure 2*. Scatter plot of average of absolute bias by simulation condition

To further investigate the item parameter recovery capability of GLIMMIX, zeta statistics were calculated to test estimation bias. The mean, standard deviation, maximum, and minimum of were displayed in Table 1, showing significant variations. For example, the standard deviations of the zeta statistic ranged from 4.65 to 85.65, which clearly suggest that the item parameter estimates produced by GLIMMIX procedure were biased.

*Table 1*. Mean, Standard Deviation, Maximum, and Minimum of the Zeta Statistic

| Length_Sample | Mean | SD | Maximum | Minimum |
|---|---|---|---|---|
| 10_100 | -0.56 | 19.09 | 25.24 | -24.08 |
| 10_200 | 0.32 | 28.96 | 36.88 | -36.15 |
| 10_400 | -0.19 | 39.39 | 48.12 | -48.05 |
| 10_800 | -0.05 | 53.60 | 66.30 | -65.88 |
| 10_1000 | -0.23 | 60.60 | 72.37 | -75.19 |
| 10_1500 | -0.22 | 74.07 | 89.10 | -90.14 |
| 10_2000 | -0.02 | 85.65 | 106.12 | -101.24 |
| 20_100 | 0.05 | 10.68 | 13.73 | -14.30 |
| 20_200 | -0.25 | 14.16 | 19.08 | -18.72 |
| 20_400 | -0.05 | 18.55 | 24.77 | -23.98 |
| 20_800 | -0.26 | 26.27 | 34.63 | -35.99 |
| 20_1000 | 0.26 | 28.38 | 35.90 | -36.29 |
| 20_1500 | -0.15 | 35.42 | 45.62 | -47.54 |
| 20_2000 | -0.02 | 40.18 | 51.90 | -48.98 |
| 40_100 | -0.11 | 5.92 | 8.96 | -9.51 |
| 40_200 | -0.05 | 7.69 | 10.41 | -11.61 |
| 40_400 | 0.10 | 9.24 | 12.56 | -13.23 |
| 40_800 | -0.03 | 12.65 | 17.07 | -17.87 |
| 40_1000 | 0.03 | 13.66 | 18.24 | -18.99 |
| 40_1500 | 0.04 | 16.85 | 22.92 | -22.95 |
| 40_2000 | 0.09 | 19.53 | 26.62 | -26.06 |
| 60_100 | 0.11 | 4.99 | 7.89 | -8.18 |
| 60_200 | 0.01 | 5.45 | 8.07 | -8.73 |
| 60_400 | 0.05 | 6.60 | 11.24 | -11.07 |
| 60_800 | 0.03 | 8.59 | 11.48 | -12.57 |
| 60_1000 | -0.03 | 9.39 | 13.73 | -13.93 |
| 60_1500 | -0.02 | 11.21 | 15.40 | -15.92 |
| 60_2000 | 0.03 | 13.07 | 18.44 | -19.16 |
| 80_100 | 0.04 | 4.45 | 8.37 | -7.70 |
| 80_200 | -0.02 | 4.53 | 7.32 | -7.83 |
| 80_400 | -0.04 | 5.32 | 8.62 | -7.87 |
| 80_800 | -0.06 | 6.62 | 9.91 | -9.83 |
| 80_1000 | 0.01 | 7.10 | 10.22 | -10.58 |
| 80_1500 | 0.01 | 8.72 | 12.87 | -13.17 |
| 80_2000 | -0.05 | 9.59 | 13.07 | -13.51 |

*Standard error estimates*. The ratios of average error variance estimates over sampling variances were plotted against the true item parameters in Figure 3. The variation in the values of these ratios was so large that those greater than 20 had to be replaced with 20 in order to have this figure displayed meaningfully. The left graph clearly indicate that shorter tests (10 and 20 items) contained the most fluctuation and thus the most bias in their standard

error estimates. The right graph shows that the shorter tests produced biased estimates across all test lengths. The longer tests (40 items and above) had ratios close to unity. Table 2 further reveals the magnitude of the variation for the 10-item tests in the average ratios. For the 20-items, most of the values were close to three. Note that smaller sample sizes (100 and 200) can produce very high ratios, for example, the 20-item and 200-person condition.

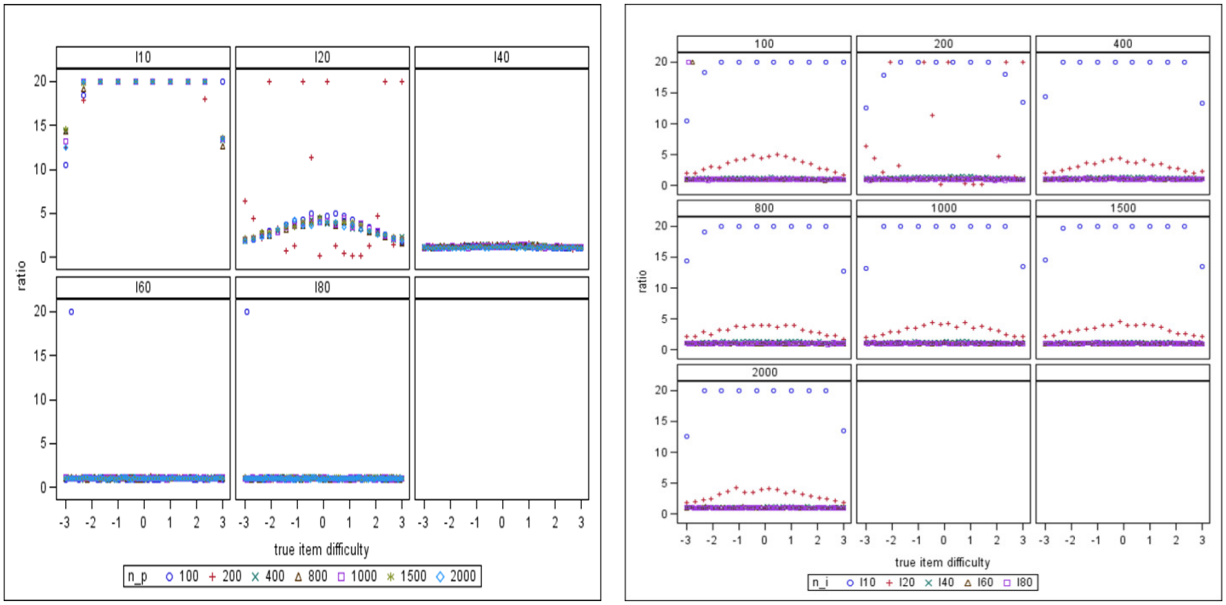


*Figure 3*. Ratios of average error variance estimates over sampling variances

*Item fit indices*. INFIT and OUTFIT mean squares are statistical indices typically used to check the model-data fit of the Rasch model. The means of both indices over a test are expected to approximate one. Figure 4 exhibits the scatter plot of INFIT mean squares against true item parameters grouped by test length and sample size respectively. The most noticeable is that with 10- and 20-item tests, the variability of this index was considerable at all sample size levels. In other words, with these short tests, the GLIMMIX procedure did not fit the Rasch model to the data very well, especially for the harder and easier items, and items in the middle of the spectrum. Moreover, the poor fit with short tests persisted as sample size increased. The results about OUTFIT mean squares are similar so their graphs are not presented here.

The mean square indices can be transformed into *t* statistics that follow roughly the standard normal distribution and thus enable significance testing.

$$t_i = \left(u_i^{1/3} - 1\right)\left(3/s_i\right) + s_i/3$$

where $u_i$ is the mean square statistic and $s_i$ is its variance. Figure 5 shows the distribution of the transformed OUTIFT mean square statistics grouped by test length and sample size respectively. Most values approach zero, which was expected. Remarkably, the 10-item test led to more variation at both ends of item difficulty spectrum. Also, sample sizes had almost no effect on this index. The *t*-transformed INFIT mean square statistics exhibited a similar pattern and thus their graphs are not presented here.

*Table 2*. Mean, Standard Deviation, Maximum, and Minimum of the Error Variance Ratio

| Length_Sample | Mean | SD | Maximum | Minimum |
|---|---|---|---|---|
| 10_100 | 34.98 | 26.54 | 105.57 | 10.45 |
| 10_200 | 26.19 | 10.26 | 40.57 | 12.51 |
| 10_400 | 28.51 | 10.91 | 44.04 | 13.32 |
| 10_800 | 28.57 | 11.20 | 44.67 | 12.69 |
| 10_1000 | 28.95 | 11.44 | 47.46 | 13.13 |
| 10_1500 | 28.83 | 11.01 | 41.59 | 13.55 |
| 10_2000 | 30.00 | 12.44 | 47.03 | 12.53 |
| 20_100 | 3.44 | 1.07 | 4.98 | 1.75 |
| 20_200 | 134.34 | 541.58 | 2433.01 | 0.12 |
| 20_400 | 3.12 | 0.79 | 4.35 | 1.92 |
| 20_800 | 3.08 | 0.75 | 4.00 | 1.63 |
| 20_1000 | 3.20 | 0.80 | 4.40 | 2.03 |
| 20_1500 | 3.23 | 0.76 | 4.60 | 2.09 |
| 20_2000 | 3.05 | 0.78 | 4.31 | 1.83 |
| 40_100 | 1.13 | 0.08 | 1.34 | 0.98 |
| 40_200 | 1.27 | 0.14 | 1.54 | 0.93 |
| 40_400 | 1.21 | 0.10 | 1.42 | 1.02 |
| 40_800 | 1.24 | 0.11 | 1.43 | 1.07 |
| 40_1000 | 1.18 | 0.09 | 1.37 | 1.01 |
| 40_1500 | 1.13 | 0.08 | 1.29 | 0.94 |
| 40_2000 | 1.07 | 0.07 | 1.23 | 0.96 |
| 60_100 | 2.27 | 10.07 | 78.99 | 0.81 |
| 60_200 | 1.05 | 0.07 | 1.23 | 0.94 |
| 60_400 | 1.04 | 0.07 | 1.20 | 0.89 |
| 60_800 | 1.01 | 0.07 | 1.18 | 0.86 |
| 60_1000 | 1.04 | 0.07 | 1.20 | 0.88 |
| 60_1500 | 1.01 | 0.06 | 1.19 | 0.91 |
| 60_2000 | 1.01 | 0.07 | 1.19 | 0.87 |
| 80_100 | 2.17 | 10.68 | 96.50 | 0.80 |
| 80_200 | 0.99 | 0.08 | 1.18 | 0.80 |
| 80_400 | 1.03 | 0.07 | 1.17 | 0.90 |
| 80_800 | 1.01 | 0.07 | 1.21 | 0.83 |
| 80_1000 | 1.01 | 0.07 | 1.17 | 0.86 |
| 80_1500 | 1.01 | 0.06 | 1.19 | 0.91 |
| 80_2000 | 0.99 | 0.06 | 1.12 | 0.87 |

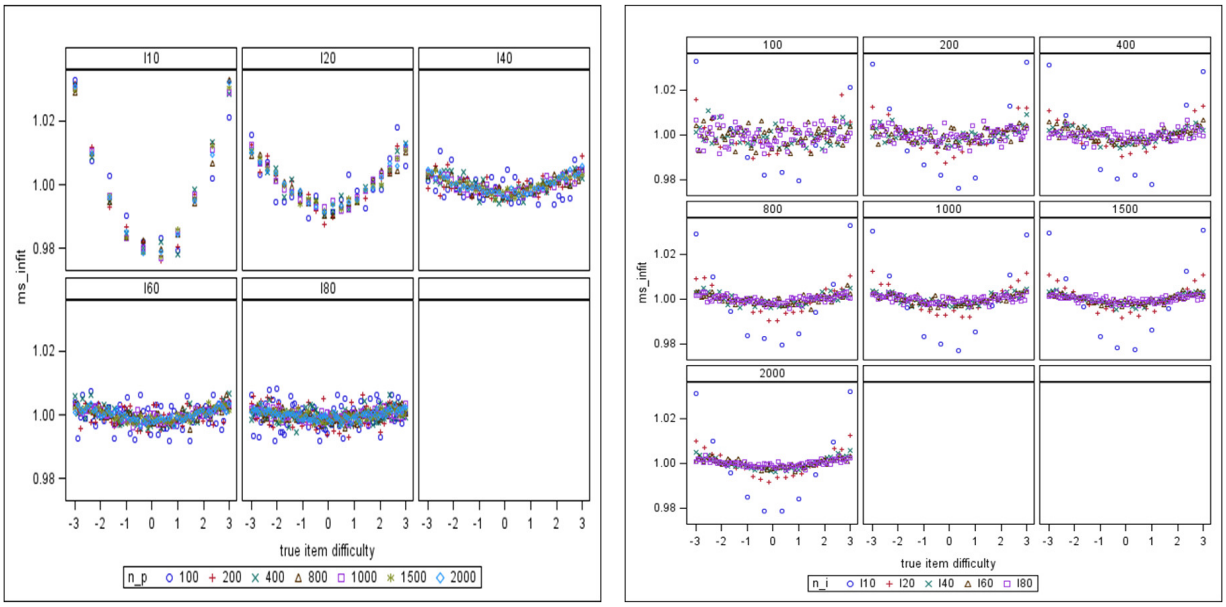


*Figure 4.* Scatter plots of INFIT mean squares against true item parameters

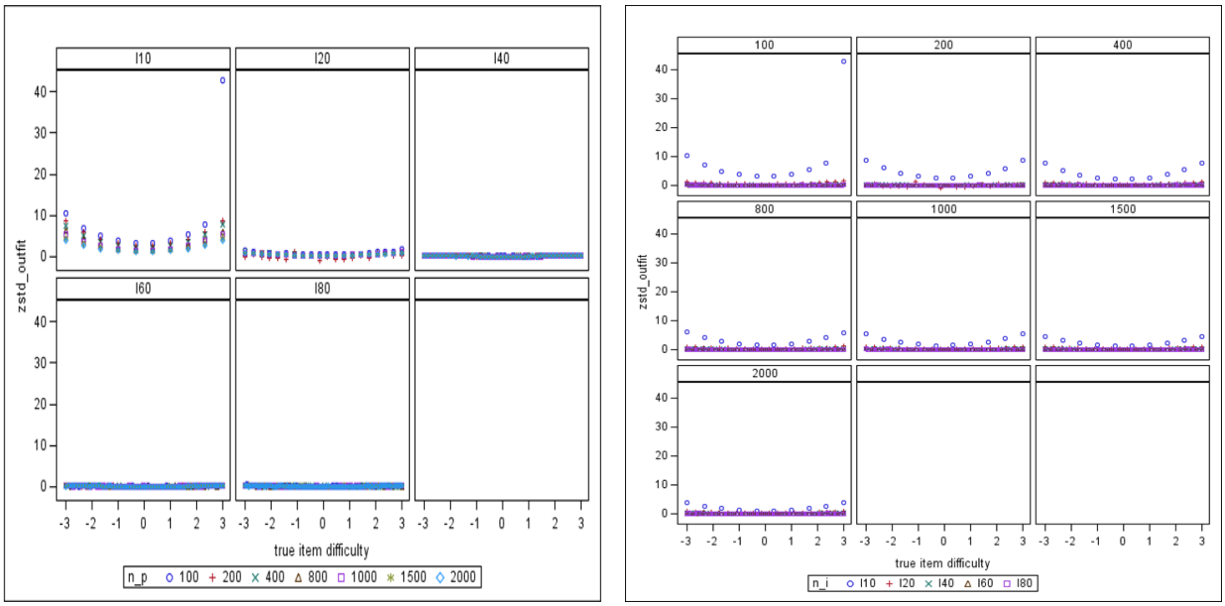


*Figure 5.* Scatter plots of transformed OUTFIT mean square index against true item parameters

The means of all four fit indices, INFIT and OUTFIT mean squares and their *t*-transformed counterparts were calculated for every simulation condition and plotted in Figure 6. The averages of the two mean squares were very close to unity across all simulation conditions, including the shortest tests. The averages of transformed OUTFIT mean squares showed much variation with the 10-item test and some variation with the 20 item variation. Figure 7 provided the scatter plot of the standard deviations of these four indices. Again, the 10-item test led to large variation in *t*-transformed OUTFIT mean squares, although for the rest of the simulation conditions these values were close to zero as expected theoretically.

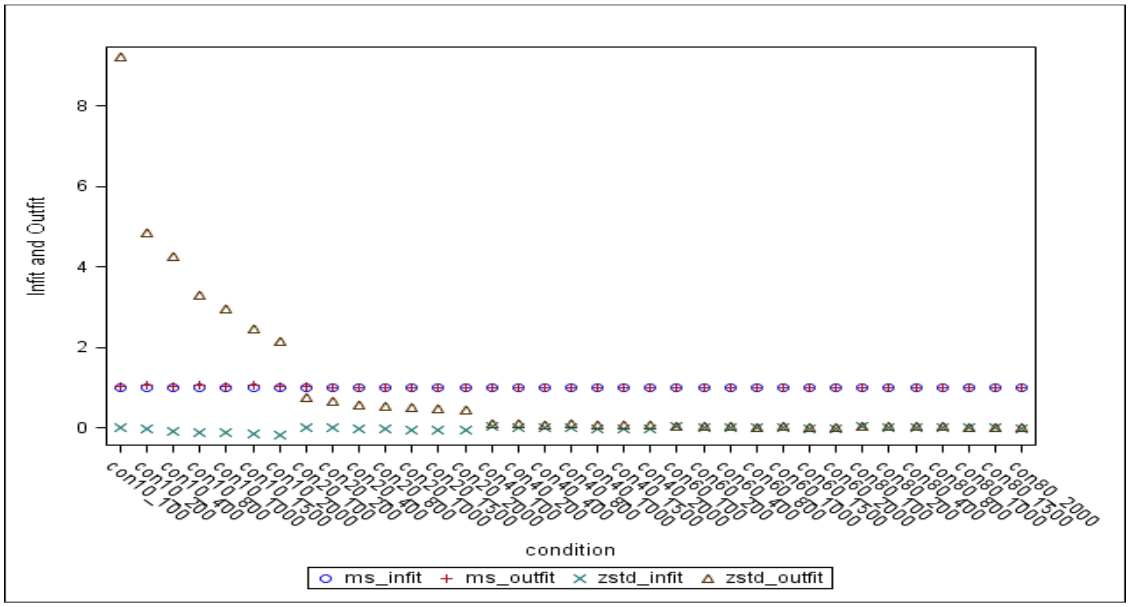


*Figure 6.* Scatter plot of average item fit indices by simulation condition

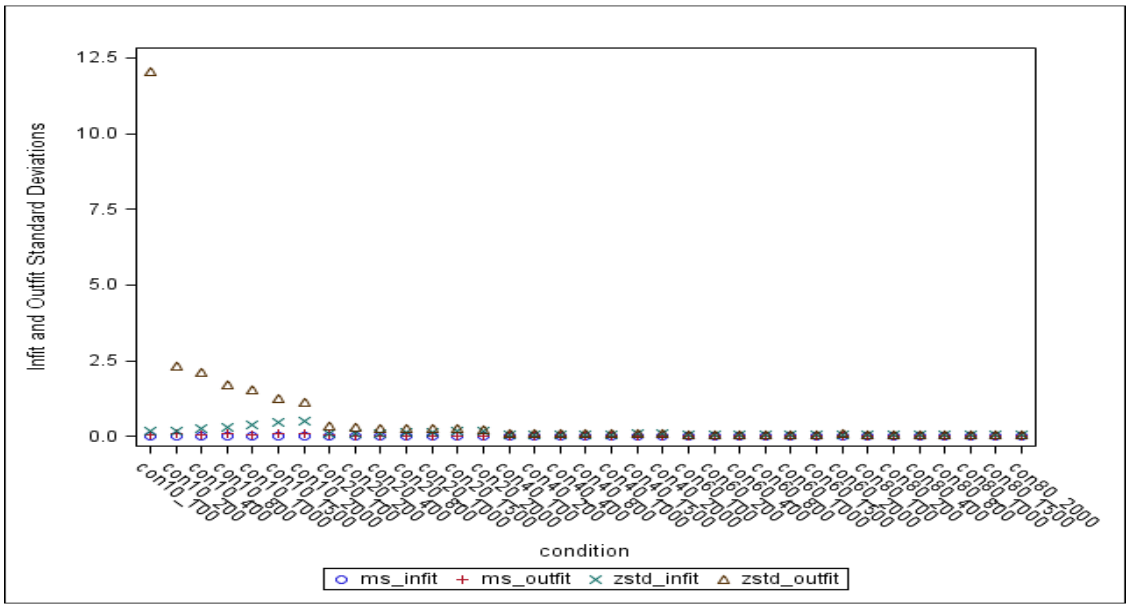


*Figure 7.* Scatter plot of average standard deviations of item fit indices by simulation condition

## CONCLUSION

Overall, the SAS GLIMMIX procedure provided respectably accurate and consistent estimation for the Rasch model, which was evident in terms of item parameter recovery, standard error estimation, and item fit indices. A word of caution is that short tests (with 20 items or fewer) can cause considerable estimation bias, larger standard error estimates, and very poor item fit. In this simulation study, different sample sizes did not have significant effect on the performance of the GLIMMIX procedure. However, the combination of short tests and small samples (200 or less) can lead to biased estimates and poor item fit. In addition, the GLIMMIX procedure can be time-consuming on a typical desktop computer when analyzing longer tests and large samples. Future research on this SAS procedure could focus on its application to more complex Item Response Theory models, different simulation conditions, and comparing its performance against specialized IRT software on shorter tests and smaller samples.

## REFERENCES

Chen, Y-H, Li, I., Kromrey, J.D. (2013). GLIMMIX_RASCH: A SAS® Macro for Fitting the Dichotomous Rasch Model. *Proceedings of the Annual South East SAS Users Group Conference*, Cary, NC: SAS Institute Inc.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Sheu, C.-F., Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. Behavior Research Methods, 37, 202-218.

Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. Psychological Methods, 8, 185-205.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Medical Care, 38(suppl II), II-28–II-42.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Wang, W. & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. Educational and Psychological Measurement, 65, 376-404.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the first author at:

Isaac Li
University of South Florida
4202 East Fowler Ave., EDU 105
Tampa, FL 33620
Fax: (813) 974-4495
Email: liy1@mail.usf.edu