SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Multiple-Group Calibration in SAS®: PROC IRT and SAS/IML®

USERS PROGRAM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Multiple-Group Calibration in SAS®: PROC IRT and SAS/IML®

Kyungyong Kim[1], Seohee Park[1], Jinah Choi[1], and Hongwook Seo[2]

[1]University of Iowa and [2]ACT

ABSTRACT

- Item response theory (IRT) has been gaining popularity in the field of educational and psychological measurement.
- IRT assumes that the probability of a correct response to an item is a function of person and item parameters.
- In IRT, many applications require a common ability scale. However, the origin and unit of measurement of the ability scale are undetermined.
- IRT estimation programs typically choose the ability scale so that mean and SD of the person parameters are 0 and 1 for the group at hand.
- When estimating parameters for two different test forms with nonequivalent groups, this common procedure yields parameter estimates that are on two different ability scales.
- Multiple-group calibration, which is supported by the IRT procedure in SAS/STAT® (SAS Institute Inc., 2013), is one approach that is often used to handle this issue.
- To conduct multiple-group calibration, the two test forms must share some common items.
- The purpose of this paper is to compare the performance of PROC IRT, a multiple-group calibration program written using SAS/IML®, and a commercial software flexMIRT® (Cai, 2013) in terms of the recovery of item parameters.

METHODS

- A simulation study was conducted to compare the performance of PROC IRT, SAS/IML®, and flexMIRT® using the two parameter logistic (2PL) model (Birnbaum, 1968) :

$$P_{ij}(\theta_i|a_j, b_j) = 1/(1 + \exp[1.7a_j(\theta_i - b_j)]),$$

  where a_j and b_j are the discrimination and difficulty parameters for item j, and θ_i is the ability parameter for person i.
- The study factors for the simulation study were as follows:

Study Factor	Form Y / Group 1 (Reference)	Form X / Group 2
Test Length	40 unique items + 20 common items	40 unique items + 20 common items
Sample Size	3,000	3,000
Ability Distribution	N(0, 1)	N(0.5, 1)

- The recovery of the item parameters were evaluated using three statistics: bias, standard error (SE), and root mean squared error (RMSE).

RESULTS

Table 1. Program comparison

Program	SAS® PROC IRT	SAS/IML®	flexMIRT®
Estimation Method	Marginal MLE	Marginal MLE	Marginal MLE
Numerical Integration	Adaptive Gauss-Hermite	Composite midpoint	Composite midpoint
Ability Distribution (Both Groups)	Normal	Empirical	Empirical
Optimization Method	Quasi-Newton	Newton-Raphson	Newton-Raphson

Note: For SAS/IML® and flexMIRT®, the distributions of ability for both groups are estimated empirically without assuming any shapes.

Table 2. Simulation results

Par.	Program	Form Y			Form X		
		Bias	SE	RMSE	Bias	SE	RMSE
a	SAS® PROC IRT	.027	.047	.055	.026	.045	.053
	SAS/IML®	.002	.048	.048	.000	.048	.048
	flexMIRT®	-.012	.047	.049	-.016	.047	.050
b	SAS® PROC IRT	-.225	.046	.232	-.260	.045	.266
	SAS/IML®	-.005	.050	.050	-.007	.051	.051
	flexMIRT®	.009	.049	.052	.007	.050	.054

- Overall, the values of bias for the item parameter estimates obtained with SAS® PROC IRT tended to be larger than those obtained with SAS/IML® and flexMIRT®.
- This tendency was more noticeable for the b-parameters than the a-parameters.
- As a result, SAS® PROC IRT yielded item parameter estimates with the largest values of RMSE.
- Item parameter estimates obtained with SAS/IML® and flexMIRT® were comparable in terms of all three evaluation criteria.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Multiple-Group Calibration in SAS®: PROC IRT and SAS/IML®

Kyungyong Kim[1], Seohee Park[1], Jinah Choi[1], and Hongwook Seo[2]

[1]University of Iowa and [2]ACT

RESULTS CONTINUED

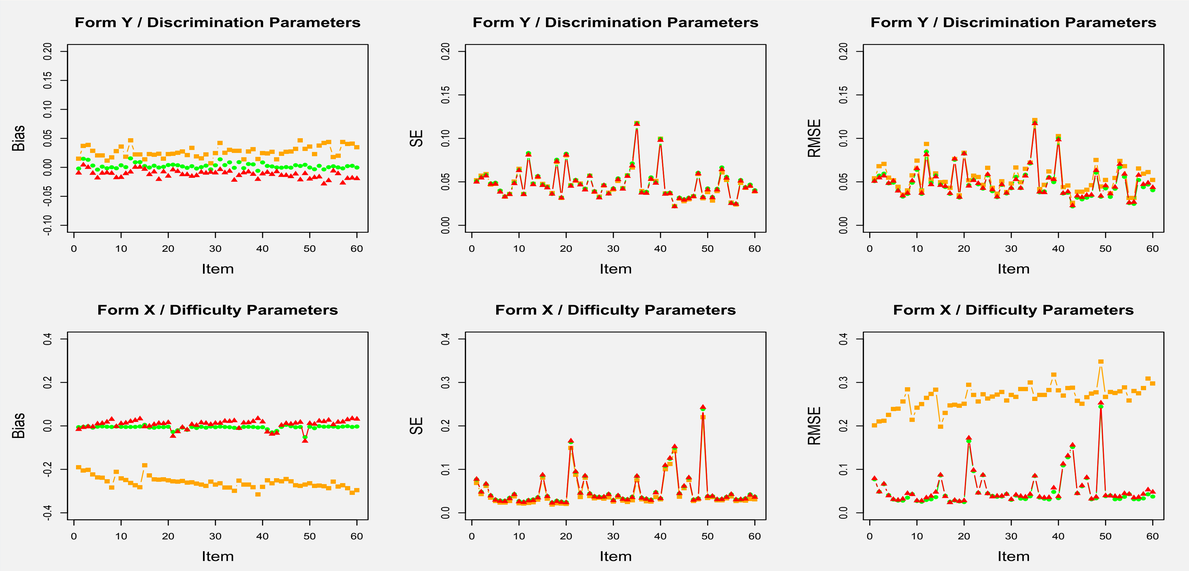

Figure 1. Conditional bias, SE, and RMSE values for the a- and b-parameters.

Note: In all figures, the results for SAS® PROC IRT are depicted with orange, the results for SAS/IML® are depicted with green, and the results for flexMIRT are depicted with red.

- SAS® PROC IRT systematically overestimates the a-parameters and underestimates the b-parameters for all items in Forms X and Y.
- Between the two item parameters, SAS® PROC IRT produced significantly less accurate b-parameter estimates in terms of both the conditional bias and RMSE statistics.
- SAS/IML® and flexMIRT yielded nearly unbiased estimates for both the a- and b-parameters.

CONCLUSIONS / LIMITATIONS

- The performance of SAS/IML® and flexMIRT are very similar in terms of the recovery of a-parameters.
- SAS® PROC IRT yields inaccurate b-parameter estimates.
  - ➔ For multiple-group calibration, the main difference between SAS® PROC IRT and the other two programs is the specification of the ability distributions during the estimation process. SAS® PROC IRT assumes abilities for both groups follow a normal distribution, whereas SAS/IML® and flexMIRT® estimate the ability distributions concurrently with the item parameters.
- SAS® PROC IRT requires significantly more computation time than the other two programs.

Program	SAS® PROC IRT	SAS/IML®	flexMIRT
Computation Time	3 minutes	0.75 seconds	0.84 seconds

Note: The SAS/IML® program written for this study only provides the item parameter estimates as the final output.

- In most IRT estimation programs, prior distributions can be assumed for the item parameters to guarantee convergence (i.e., marginalized Bayesian estimation; Mislevy, 1986). However, SAS® PROC IRT only supports the marginal maximum likelihood estimation method (Bock and Aitkin, 1981).
- Because of the small number of conditions examined in the simulation study, this study is limited in its ability to generalize the findings to measurement conditions that are not included in the simulation study.

REFERENCES

Birnbaum, A. (1968). Estimation of an ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 423-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Cai, L. (2017). flexMIRT® version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51*, 177-195.

SAS Institute Inc. (2013). SAS/STAT® 13.1 User's Guide: The IRT procedure [Computer software]. Cary, NC.

Contact Information: kyungyong-kim@uiowa.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.