SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Evaluation of PROC IRT* Procedure for Item Response Modeling

*Available in SAS/STAT® 13.1, 13.2 & 14.1

Evaluation of PROC IRT Procedure for Item Response Modeling

Yi-Fang Wu

Measurement Research, ACT, Inc.

ABSTRACT

- In the SAS/STAT®13.1, 13.2 and 14.1, the **PROC IRT** procedure enables item response modeling and latent trait (e.g., ability) estimation for various item response theory (IRT) models
- Under a wide-spectrum of educational and psychological research, IRT gains popularity in literature and in practice
- As a technical improvement, **PROC IRT** offers a great choice to the growing population of IRT users
- **PROC IRT** supports several item response models for binary responses like the one-, two-, three-, and four-parameter models and the response models for ordinary responses such as the graded response models with a logistic or probit link (An & Yung, 2014)
- Considering common testing conditions (Anastasi, & Urbina, 1997), this paper intended to evaluate the performance of **PROC IRT** in terms of ***item parameter recovery***
- IRT models for dichotomous response data were investigated: the one-parameter logistic (1PL) (a.k.a. Rasch model; Rasch, 1960), the two-parameter logistic (2PL) (Birnbaum, 1968), and the three-parameter logistic (3PL) (Birnbaum, 1968; Lord, 1980) models
- The ***pros and cons*** of **PROC IRT** against **BILOG-MG** 3.0 (Zimowski, Muraki, Mislevy, & Bock, 2003) were presented
- For practitioners of IRT models, the development of the IRT-related analysis in SAS should be inspiring

METHODS

- The 3PL model has the generic form : $P_{ij}(\theta_j | a_i, b_i, c_i) = c_i + (1- c_i)/(1 + \exp[-Da_i (\theta_j - b_i)])$, where $P_{ij}(\theta_j | a_i, b_i, c_i)$ is the probability that the examinee j with a θ_j ability answers the item i correctly; a_i, b_i and c_i denote item discrimination, item difficulty, and pseudo-guessing parameters, respectively. D equal to 1.702 is the scaling constant. Letting $c_i = 0$ for all items results in the 2PL model; finally, letting $c_i = 0$ and $a_i = 1$ for all items results in the 1PL model
- The 3PL model allows each item to vary in the item difficulty, discrimination, and pseudo-guessing parameters; the 1PL model is the most constraint and the 2PL model is in-between
- In simulations, factors and levels under investigation were in Table 1; for each condition, 100 replications were done

Table 1. Factors and Levels of Interest

Factor	Description
Model	1PL, 2PL & 3PL
Sample Size	250 (small) & 1000 (large) examinees
Test Length	20 (short) & 40 (medium) items within a test
Underlying Ability Distribution	Normal (Nor), negatively-skewed (Neg) & positively-skewed (Pos)
Test Composition	Tests with moderate average item difficulty with moderate average item discrimination (TC1), hard tests with moderate average item discrimination (TC2) & easy test with low average item discrimination (TC3)

Tale 2. True Parameter Distributions

Parameter	Distribution	Test Composition
Discrimination (a)	*Beta4*(5, 5, 0.1, 2)	TC1
	Beta4(6, 2, 0.1, 2)	TC2
	Beta4(2, 6, 0.1, 2)	TC3
Difficulty (b)	*Beta4*(5, 5, –3, 3)	TC1-TC3
Pseudo-Guessing (c)	Uniform(0, 0.25)	TC1-TC3
Ability	N(0, 1)	Nor
	Gamma(10, 1.5) rescaled so that mean = 0 & variance = 1	Pos & Neg* (*mirrored)

RESULTS AND DISCUSSION

Table 2. Correlations (Aggregated over Replications) between True Item Parameters and Estimates from PROC IRT and BILOG-MG

			Test Length		Sample Size		Ability Distribution			Test Characteristics		
Model	**Parameter**	**Procedure**	**Short**	**Medium**	**Small**	**Large**	**Nor**	**Neg**	**Pos**	**TC1**	**TC2**	**TC3**
3PL	Discrimination (a)	Proc IRT (P)	.464	.555	.460	.559	.560	.392	.577	.380	.371	.777
		BILOG-MG (B)	*.704*	*.754*	*.634*	*.823*	*.766*	*.651*	*.770*	*.712*	*.588*	*.886*
		Corr(P, B)*	.794	.790	.849	.735	.785	.792	.799	.778	.681	.917
	Difficulty (b)	Proc IRT	.850	.843	.838	.855	.843	.847	.848	.958	.820	.761
		BILOG-MG	*.919*	*.926*	*.908*	*.937*	*.925*	*.916*	*.927*	*.981*	*.955*	*.831*
		Corr(P, B)	.886	.882	.876	.891	.884	.887	.880	.959	.791	.901
	Pseudo-Guessing (c)	Proc IRT	.298	.399	.312	.385	.366	.324	.356	.454	.413	.179
		BILOG-MG	*.467*	*.601*	*.483*	*.586*	*.539*	*.566*	*.498*	*.596*	*.841*	*.166*
		Corr(P, B)	.242	.356	.376	.223	.291	.246	.361	.420	.476	.002
2PL	Discrimination (a)	Proc IRT	.856	.851	.789	.918	.878	.793	.889	.851	.771	.939
		BILOG-MG	*.883*	*.877*	*.820*	*.940*	*.895*	*.845*	*.899*	*.874*	*.826*	*.939*
		Corr(P, B)	.982	.972	.974	.981	.979	.969	.983	.988	.946	.997
	Difficulty (b)	Proc IRT	.874	.849	.848	.876	.860	.862	.863	.957	.867	.761
		BILOG-MG	*.974*	*.968*	*.956*	*.985*	*.972*	*.967*	*.972*	*.993*	*.974*	*.944*
		Corr(P, B)	.893	.866	.869	.890	.878	.886	.876	.959	.845	.836
1PL	Difficulty (b)	Proc IRT	.992	.972	.966	.998	.995	.965	.985	.998	.963	.985
		BILOG-MG	*.995*	*.994*	*.992*	*.998*	*.995*	*.995*	*.994*	*.998*	*.992*	*.994*
		Corr(P, B)	.995	.975	.970	1	1	.968	.986	1	.968	.987

Note. Corr(P, B) denotes the correlation between the **PROC IRT** estimates and the **BILOG-MG** estimates.

Correlations:

- Overall, the averaged correlations over replications for **BILOG-MG** tended to be *higher* than the correlations for **PROC IRT**
- The agreement between the true and estimated values seemed to be the *highest* for b-parameters; furthermore, the agreement for b was the *highest* for **1PL models**
- For 3PL models, the agreement between the true and estimated values was the *lowest* for c-parameters
- The agreement between the true a-parameters and a-estimates *decreased* whenever c needed to be estimated
- a-parameters could be better estimated when *tests were easy and less discriminating* (i.e., TC3)
- The impact of test length, sample size and ability distribution on agreement seemed random. Relatively speaking, results from *normal and positively-skewed distributions* were alike

- **Evaluation criteria**: ***Correlation*** of the true and estimated parameters, ***bias*** (BIAS), ***absolute-bias*** (ABSB), and ***root mean square error*** (RMSE)

Evaluation of PROC IRT Procedure for Item Response Modeling

Yi-Fang Wu

Measurement Research, ACT, Inc.

RESULTS AND DISCUSSION CONTINUED

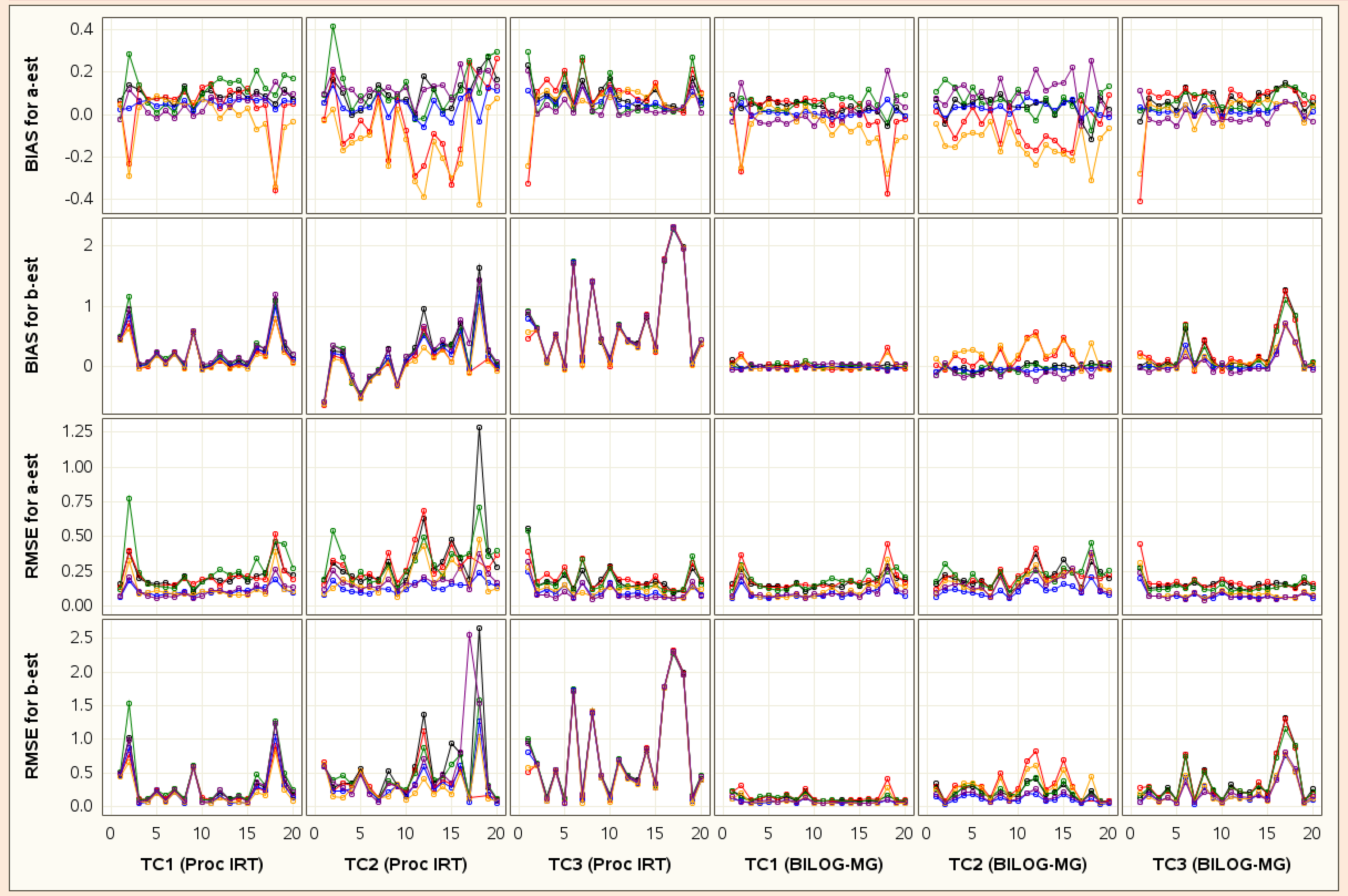

Figure 1. BIAS and RMSE of *a*- and *b*-estimates for 2PL Models (short tests)

Note. In both figures, **black** for **Nor_250**, **red** for **Neg_250**, **green** for **Pos_250**, **blue** for **Nor_1000**, **orange** for **Neg_1000** and **purple** for **Pos_1000**.

BIAS (Selective Results Shown), **ABSB** (Results Not Shown) and **RMSE** (Selective Results Shown):

- For 1PL models, **PROC IRT** had *lower* ABSB for *b*-estimates than **BILOG-MG** (.816 vs. 1.588); note that no *b*-priors was used for 1PL calibrations
- For 2PL and 3PL models, **PROC IRT** had *higher* ABSBs for *a*- and *b*-estimates than **BILOG-MG**, but had *lower* ABSB for *c*-estimates (.057 vs. .076) in 3PL models
- Peculiar values were *less likely* occur from **BILOG-MG** due to the use of prior-constraints in estimation; thus, the BIAS and RMSE were *smaller* for **BILOG-MG**
- Difference in sample size had *minor* impact; difference in ability distribution mostly affected *a*-parameter estimation (and sometimes *c*-parameter estimation), especially when tests were just moderately discriminating (i.e., TC1 & TC2), *negatively-skewed* distributions could result in negative bias (i.e., underestimating *a*)
- **BILOG-MG** tended to *overestimate* *c-parameters*
- *Large* RMES of *c*-estimates indicated *c*-parameter estimation was *challenging* regardless of the estimation procedure/program

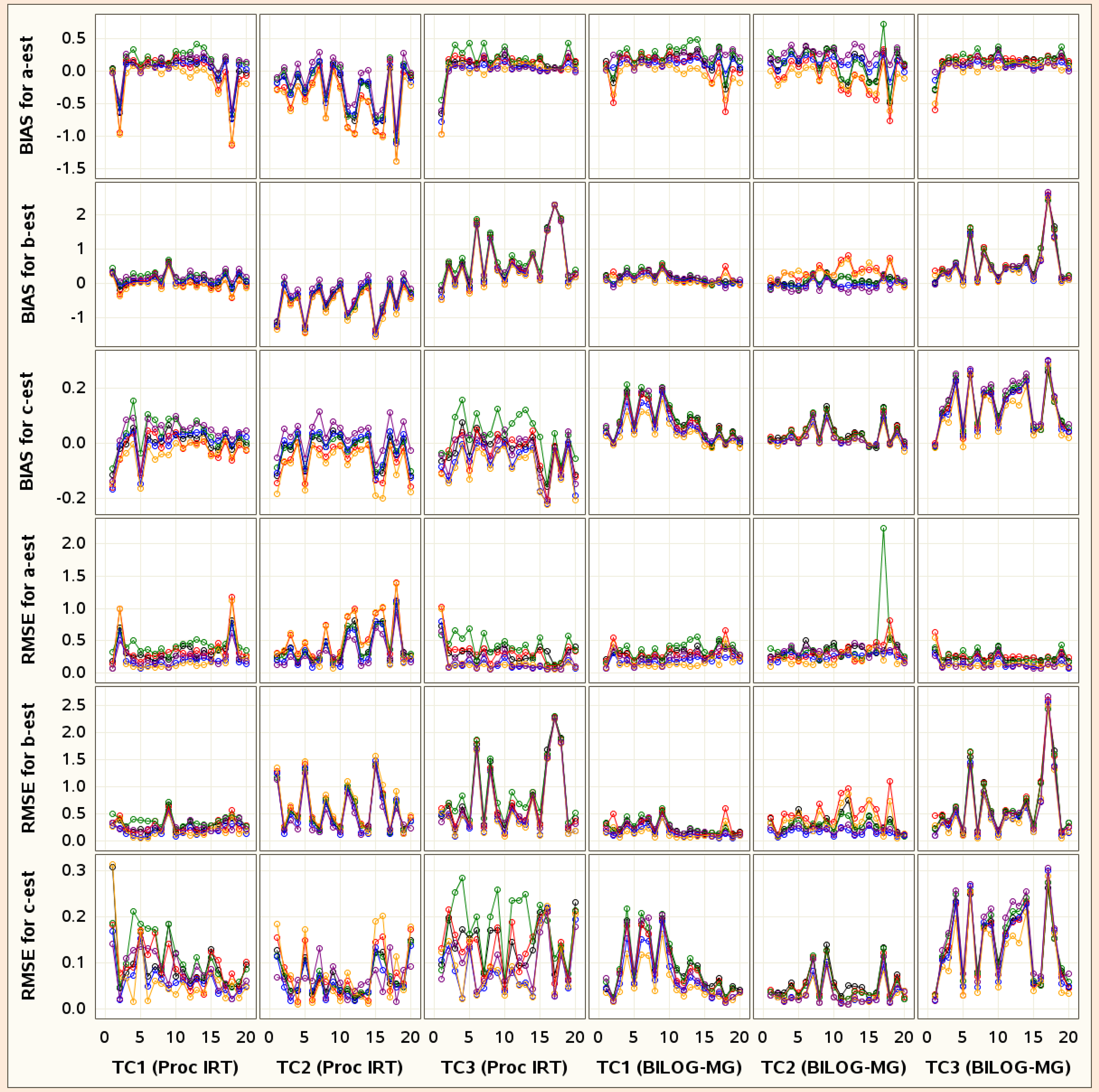

Figure 2. BIAS and RMSE of *a*-, *b*- and *c*-estimates for 3PL Models (short tests)

COMPARISON

Procedure / Feature	PROC IRT	BILOG-MG
Dimensionality	Both ***unidimensional*** and ***multidimensional*** IRT models are considered (*for the latter, the parameter accuracy and practical feasibility needs further investigation*)	Only *unidimensional* IRT models are considered
Item Parameter Recovery	Good for 1PL models	Good for 1PL models; ***better*** for 2PL and 3PL models
Estimation Convergence	Algorithm can be ***converged*** most of time; no solution for response level less than 2	More likely to fail for small datasets under *3PL* models unless ***priors*** are requested
Estimation Options	Latent trait score (factor score) estimation is available, including *maximum likelihood*, *maximum a posteriori* and *expected a posteriori*	Same as left, in addition that ***priors*** of item parameter estimation can be requested to prevent peculiar values
Response Data Types	***Dichotomous*** and ***polytomous*** response items can both be estimated	Only *dichotomous* responses are acceptable
Computing Feasibility	Under 3PL models, *one* dataset of 40-item and 1000-examines can be calibrated within 1 minute	Under the same condition, ***ten*** datasets of the same size can be calibrated within 1 minute
Others	**Multi-group estimation** and **item and test characteristic curves** are available upon request	Same as left

REFERENCES

- **An, X., & Yung, Y.-F.** (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *Proceedings of the SAS® Global Forum 2014 Conference*, Paper 364. Cary, NC: SAS Institute.
- **Anastasi, A., & Urbina, S.** (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.
- **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.
- **Rasch, G.** (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institute.
- **Lord, F. M.** (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.
- **SAS Institute Inc.** (2014). The IRT procedure (experimental). *SAS/STAT® 13.2 User's Guide*. Cary, NC: (author). Retrieved from http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_irt_toc.htm
- **Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003)**. BILOG-MG: Multiple group IRT analysis and test maintenance for binary items (Version 3.0) [Compute program] Lincolnwood, IL: Scientific Software International.

For More Information

Please contact Yi-Fang Wu at Yi-Fang.Wu@act.org

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF