American Journal of Educational Research, 2019, Vol. 7, No. 11, 865-871
Available online at http://pubs.sciepub.com/education/7/11/17
Published by Science and Education Publishing
DOI:10.12691/education-7-11-17

A Comparison between BMIRT and IRTPRO: A Simulation Study of a Multidimensional Item Response Model

Tingxuan Li*

Graduate School of Education, Shanghai Jiao Tong University, Shanghai, China
*Corresponding author: litingxuanpurdue@gmail.com

Received October 10, 2019; Revised November 13, 2019; Accepted November 25, 2019

Abstract The objective of this study is to provide comparative information on two software programs- *IRTPRO version 2.1 for Windows* and *BMIRT*. In educational measurement, software programs are being developed and updated rapidly. By using a small-scale simulation study on a *two-parameter logistic model* in multidimensional item response theory, this study is to examine the bias values and root mean square error values produced by both programs. Other than item parameter recovery, the comparisons about run time and user interface were also made. The results showed that BMIRT was better in estimating item slope parameters. However, in terms of run time, it is much slower than IRTPRO. In addition, IRTPRO's interface is much more user friendly than BMIRT's. Screenshots of conducting item calibrations for both programs are in Appendix A.

Keywords: *multidimensional item response theory, educational measurement, simulation*

Cite This Article: Tingxuan Li, "A Comparison between BMIRT and IRTPRO: A Simulation Study of a Multidimensional Item Response Model." *American Journal of Educational Research*, vol. 7, no. 11 (2019): 865-871. doi: 10.12691/education-7-11-17.

1. Introduction

In item response theory (IRT), multiple software programs, for example, PARSCALE [1] and MULTILOG [2], are available for psychometric analyses, research, and score reporting. Studies related to software programs often appear on journals in the format of a short review or software note [3,4]. The content of such studies consists of information on estimation methods, the associated code, and/or screenshots taken from the program interface. Paek and Han [4] wrote a review for IRTPRO version 2.1 for Windows (developed by [5]). The review summarized the user guide of IRTPRO program. It also highlighted the parameterization in the nominal response models. The default or commonly being used estimation method is Bock-Aitkin expectation-maximization algorithm (BA-EM; [6]). Schwarz [3] wrote a review for Bayesian Multivariate IRT Toolkit (BMIRT, developed by [7]). BMIRT contains a set of programs: LinkMIRT on linking program, BMIRT II on item calibration, SimuMIRT and SimuMCAT on simulation for multidimensional test forms. The commonly used program is BMIRT II, which adopts Markov chain Monte Carlo (MCMC) algorithm to estimate parameters. Schwarz [3] further pointed out that more detailed comparisons among IRTPRO, EQSIRT [8], or BMIRT were needed. The results from the comparisons may benefit the practitioners who employ multidimensional item response theory (MIRT) models.

2. Research Goal

MIRT models have become increasingly popular in the last decade [9,10]. This study is to expand the literature on MIRT software by examining the following details: (a) parameter specification, (b) screenshots to show step-by-step procedure for item calibrations, (c) comparisons on parameter recovery, as well as (d) the additional aspects (e.g., run time and cost). The goal is to assist practitioners in the use of software programs by looking comparative information on IRTPRO and BMIRT. The item calibration comparison is made based on data simulated from a MIRT model. In a broader sense, the software comparison is also the algorithms comparison, that is, the BA-EM algorithm and the MCMC algorithm.

Practitioners may consider a MIRT model in the following contexts. First, the theory explicitly points out that more than one latent ability is required. For example, a writing test is designed to assess two competency areas: *Write for different purposes* along with *Write using conventions*. Second, the assumption of unidimensional IRT model is violated. The practitioners may want to delete items, to assume model robustness, or to form the subscales separately with IRT models [11,12]. When MIRT modeling is used, the primary advantage is that item parameters and latent correlated ability parameters are simultaneously produced in the estimation process, which in turn can enhance the model-data fit [13,14]. Two

major schools in MIRT analysis are *compensatory models* (CM, [15,16] and *non-compensatory models* (NCM, [17,18]. When using CM, the latent abilities can interact. For example, a test-taker who is low on one latent continuum can be "saved" by being high on the others. For NCM, the test-taker must demonstrate competency for all latent abilities in order to correctly answer an item. Further, NCM has estimation challenges. Thus, CM is the focus of this study.

3. Model

3.1. Model Choice

MIRT models are considered by a variety of practitioners working on competency assessment, patient-reported outcome measurements, and psychological inventory [19,20]. In this study, the main focus is only on how MIRT is used in competency assessment. Because both IRTPRO and BMIRT can handle item calibrations for a wide range of MIRT models, the initial step of this study was to determine which model or models should be examined. The existing literature is used to identify which models appear often when handling real-world datasets in competency assessment. This systematic review started from two journals - (1) Applied Psychological Measurement, (2) Educational Measurement: Issues and Practice from the years 2000 to 2013. While many journals publish articles focusing on the use of MIRT in competency assessment, the journals used here are representative across the publication pool.

Two articles are found that had item calibrations using real-world data to support the usage of MIRT. Table 1 shows the details about the articles. Other research studies either focused on item calibrations in MIRT *without* real-world applications or focused on other aspects of the psychometric properties in MIRT (e.g., differential item functioning procedure in MIRT). Both articles used *between-item two-dimensional two-parameter logistic model*. The term, *between-item*, simply indicates each item measures only one latent ability [21]. Thus, in the simulation study below, a *between-item two-dimensional two-parameter logistic model* is used.

Table 1. Two Real-world Data for Item Calibration

Article	Test (with specific latent abilities)	Test-taker
[22]	English usage test with 31 multiple-choice items (ability to detect punctuation error & word usage error)	College freshman in the US
[23]	Math achievement test with 35 multiple-choice items (general mathematics ability & spatial ability)	9th graders in Canada

3.2. Parameters

This study uses the following notations. A matrix X contains binary response data: $X = \{X_{ij}\}$, where i is examinee, $i = 1, 2, 3, \ldots, N$ and j is item, $j = 1, 2, \ldots, J$. If D is the number of dimensions/subscales on the test, then the latent ability parameter for examinees is a D-dimension vector: $\theta_i = (\vec{\theta}_1, \vec{\theta}_2, \ldots, \vec{\theta}_N)^T$. The probability of getting a correct response on item j given the ability $\vec{\theta}_i$ in a multidimensional two-parameter logistic model is denoted as Equation 1 below. The scale difficulty parameter, that is, the intercept parameter, is $\beta_{1,j}$. The discrimination parameter, that is, the slope parameter for dichotomous item j, is $\vec{\beta}_{2,j} = (\beta_{2,1,1}, \ldots, \beta_{2,J,D})$. Note that among all item parameters in the model, only the slope parameter is represented by a vector.

$$P_{i,j} = P\left(X_{i,j} = 1 \mid \vec{\theta}_i, \vec{\beta}_j\right) = exp\left(\vec{\beta}_{2,j}\vec{\theta}_i^{\,T} + \beta_{1,j}\right) + \left[1 + exp\left(\vec{\beta}_{2,j}\vec{\theta}_i^{\,T} + \beta_{1,j}\right)\right]. \quad (1)$$

3.3. Simulated Data

The data are generated from a *between-item two-dimensional two-parameter logistic model* in R environment [24] where the number of test-takers is 1000, I = 1000; the number of items is 20, J = 20; the correlation between two latent abilities is $r = 0.5$. The latent ability parameters are drawn from a multivariate normal distribution, $\theta_i \sim$ MVN(0, Σ), where the mean and variance vector is μ = {0, 0}, σ ={1, 1}. Table 2 below is the true item parameters used in data generation. This table appears on the textbook written by M. Reckase ([16], p.126). *In order to compare the estimated parameters, the bias values and the root mean square error (RMSE) values are calculated for each set of item calibration. Equation 2 and Equation 3 show how they are computed, where R is the number of replicates, in this study, R = 20. The comparison criteria are the lower magnitudes on bias and RMSE, the better the estimation method is in terms of statistical stability and accuracy. The parameters at interests, in this study, as defined in Equation 1 which are* $\vec{\beta}_j = (\vec{\beta}_{2,j}, \beta_{1,j})$.

$$Bias\left(\hat{\beta}\right) = \frac{\sum_r^R (\hat{\beta}_r - \beta)}{R}. \quad (2)$$

$$RMSE\left(\hat{\beta}\right) = \sqrt{\frac{\sum_r^R \left(\hat{\beta}_r - \beta\right)^2}{R}}. \quad (3)$$

Table 2. True Parameters ([16], on page 126) used in the Simulation

Item ID	$\beta_{2,j,1}$	$\beta_{2,j,2}$	$\beta_{1,j}$
1	1	0	0
2	1	0	0.5
3	0.6	0	-0.9
4	0.8	0	-1.2
5	1.3	0	1
6	1.7	0	-1
7	1.1	0	-1.5
8	0.7	0	-0.6
9	0.9	0	1.2
10	2	0	0.7
11	0	1	0
12	0	0.9	0.9
13	0	1.5	2.5
14	0	1.3	1.7
15	0	1.6	-0.5
16	0	1.1	-0.8
17	0	0.8	1
18	0	1.8	0.7
19	0	0.7	0.5
20	0	2	-1

4. Results

Throughout the item calibration, the default estimation method in IRTPRO was used, which was BA-EM algorithm. For the item calibration in BMIRT software, the MCMC algorithm requires the specified iteration and burn-in values. Therefore, an iteration of 10000 and a burn-in values of 4000 were manually specified. In terms of run time, IRTPRO is notably faster than BMIRT. It took IRTPRO about 22 seconds to run each replicate and took BMIRT about 64 seconds to run each replicate. For each replicate, every item calibration estimated totally 40 parameters for 20 items. It is worth noting that both programs produced similar RMSE values across 40 parameters. In contrast, the difference between the two programs on bias values is larger. BMIRT produced smaller bias values than IRTPRO did. Table 3 is the RMSE and bias values produced by each program. In terms of bias values, for most estimated parameters, BMIRT is slightly better.

Figure 1a shows the visual gap of bias values between both programs across 20 slope parameters. The programs differ greatly on Item 13, Item 14, and Item 15. The difference of bias values is around 0.1 across these 3 items. The difference on Item 13 is the largest, nearly 0.16. Figure 1b shows the visual gap of bias values between both programs across 20 intercept parameters. The difference between the programs on intercept parameters is not as dramatic as slope parameters. The largest difference is found on Item 14 which is about 0.06.

Finally, for software program cost, BMIRT is a free software. IRTPRO is free for student version (which was used to conduct this study). Unlike the full-version software, the student version has the following maximal values for data input, for example, the total number of items can be estimated is no more than 25 items. In terms of user interface, IRTPRO is a much more user-friendly program compared with BMIRT. Appendix A at the end of this study shows the screenshots of conducting the item calibrations.

Table 3. Bias and RMSE from Each Program

	Bias on Intercept Parameter		RMSE on Intercept Parameter		Bias on Slope Parameter		RMSE on Slope Parameter	
ID	BMIRT	IRTPRO	BMIRT	IRTPRO	BMIRT	IRTPRO	BMIRT	IRTPRO
1	0.07	0.07	0.16	0.16	0.04	0.05	0.16	0.16
2	0.09	0.09	0.13	0.14	0.03	0.03	0.13	0.14
3	0.07	0.07	0.13	0.13	0.00	-0.01	0.13	0.13
4	0.09	0.08	0.14	0.14	0.01	0.02	0.14	0.14
5	0.16	0.17	0.23	0.24	0.06	0.08	0.23	0.24
6	0.16	0.16	0.21	0.21	-0.05	-0.09	0.21	0.21
7	0.08	0.06	0.15	0.15	0.05	0.09	0.15	0.15
8	0.05	0.05	0.12	0.12	0.06	0.10	0.12	0.12
9	0.11	0.13	0.20	0.22	0.06	0.12	0.20	0.22
10	0.17	0.18	0.22	0.22	0.00	-0.04	0.22	0.22
11	0.02	0.04	0.12	0.13	0.05	0.08	0.12	0.13
12	0.01	0.01	0.15	0.15	0.04	0.07	0.15	0.15
13	0.07	0.07	0.24	0.31	0.03	0.19	0.24	0.31
14	0.00	0.06	0.19	0.21	0.10	0.19	0.19	0.21
15	0.02	0.02	0.11	0.11	0.20	0.32	0.11	0.11
16	0.01	0.03	0.10	0.11	0.08	0.14	0.10	0.11
17	0.05	0.07	0.11	0.13	0.06	0.12	0.11	0.13
18	0.02	0.07	0.17	0.18	0.12	0.19	0.17	0.18
19	0.01	0.03	0.10	0.11	0.02	0.03	0.10	0.11
20	0.01	0.02	0.17	0.18	0.08	0.14	0.17	0.18

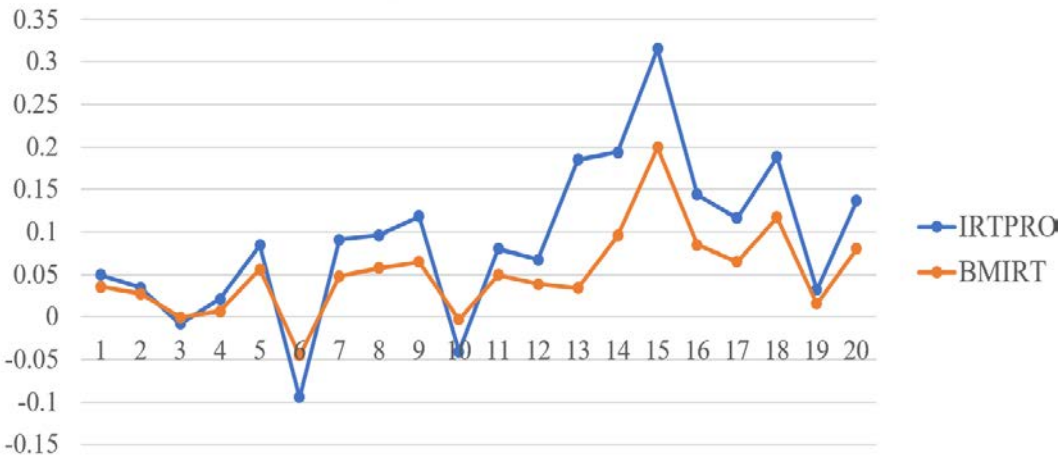


Figure 1a. Comparison between IRTPRO and BMIRT on slope parameters for bias values

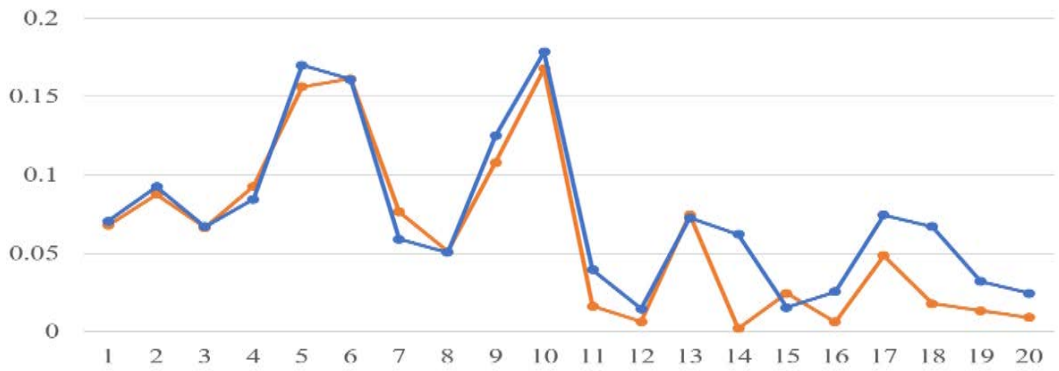


Figure 1b. Comparison on intercept parameters for bias values

5. Conclusion

In this research, a small-scale simulation study was conducted. The findings showed that BMIRT software had the advantage of producing smaller magnitude on bias, but it had longer run time. IRTPRO had greater magnitude on the bias value but it was user-friendly and had much less run time. In addition, BMIRT is free. The practitioners and researchers can obtain it without cost. In contrast, IRTPRO requires purchasing license. The limitation of this research is the lack of a variety of simulation conditions. The future simulation design should include more dimensions, for example, a *between-item three-dimensional two-parameter logistic model* should be considered. Moreover, the findings of this research are consistent with the existing literature. For example, Yavuz1and Hambleton [25] compared BMIRT program and flexMIRT program. They found that the MCMC algorithm embedded in BMIRT program was indeed much slower than the BA-EM algorithm embedded in flexMIRT program when estimating MIRT models.

6. Conclusion

In educational measurement, the software related issues have received attentions from practitioners. In this research, two software programs - BMIRT and IRTPRO 2.1 were compared in terms of item parameter recovery for a *between-item two-dimensional two-parameter logistic model*. Multidimensional item response theory (MIRT) is becoming popular because of its capacity to produce subscores. When more than one ability is measured on a test, MIRT models can estimate the latent abilities simultaneously, which in turn has decreased the standard error of estimation. The findings of this research have provided the information to practitioners when they make choice on MIRT-related software.

References

[1] Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

[2] Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

[3] Schwarz, R. (2015). A review of BMIRT TOOLKIT: BMIRT for Bayesian Multivariate IRT. *Applied Psychological Measurement, 39*(2), 155-159.

[4] Paek I. & Han K. T. (2013). IRTPRO 2.1 for Windows (item response theory for patient-reported outcomes). *Applied Psychological Measurement, 37*(3), 242-252.

[5] Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

[6] Bock, R. D., Aitkin, M. (1981). MML estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

[7] Yao, L. (2003). BMIRT: Bayesian Multivariate Item Response Theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

[8] Wu, E. J., & Bentler, P. M. (2011). EQSIRT: A user-friendly IRT program [Computer software]. Encino, CA: Multivariate Software

[9] Chen, W-H., & Thissen, D. (1997). Local dependence indices for item pairs using response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

[10] Falk, C. F., & Monroe, S. (2018). On Lagrange multiplier tests in multidimensional item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement, 78*(1), 653-678.

[11] Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*(1), 279-293.

[12] Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

[13] Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[14] Chalmers, R. P. (2016a). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1-38.

[15] Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* MA: Addison-Wesley.

[16] Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

[17] Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *40*(2), 175-186.

[18] Hsu, C-L., & Wang, W-C (2018). Multidimensional computerized adaptive testing using non-compensatory item response theory models. *Applied Psychological Measurement, 43*(6), 464-480.

[19] Park, J. Y., Cornillie, F., van der Maas, H., & Van Den Noortgate, W. (2019). A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in psychology, 10*(1), 1-10.

[20] Embretson, S. E. (2010). Measuring psychological constructs. Washington, DC: American Psychological Association.

[21] Adams, R. J., Wilson, M., & Wang, W-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

[22] Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

[23] Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

[24] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

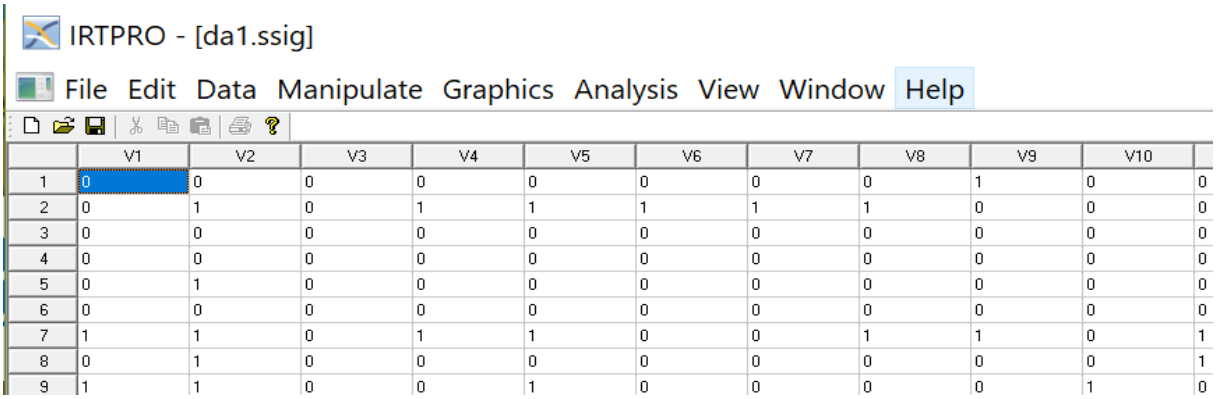
[25] Yavuz, G., & Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement, 77*(2), 263-274.

Appendix A

Screenshots—Step-by-step for item calibration

IRTPRO

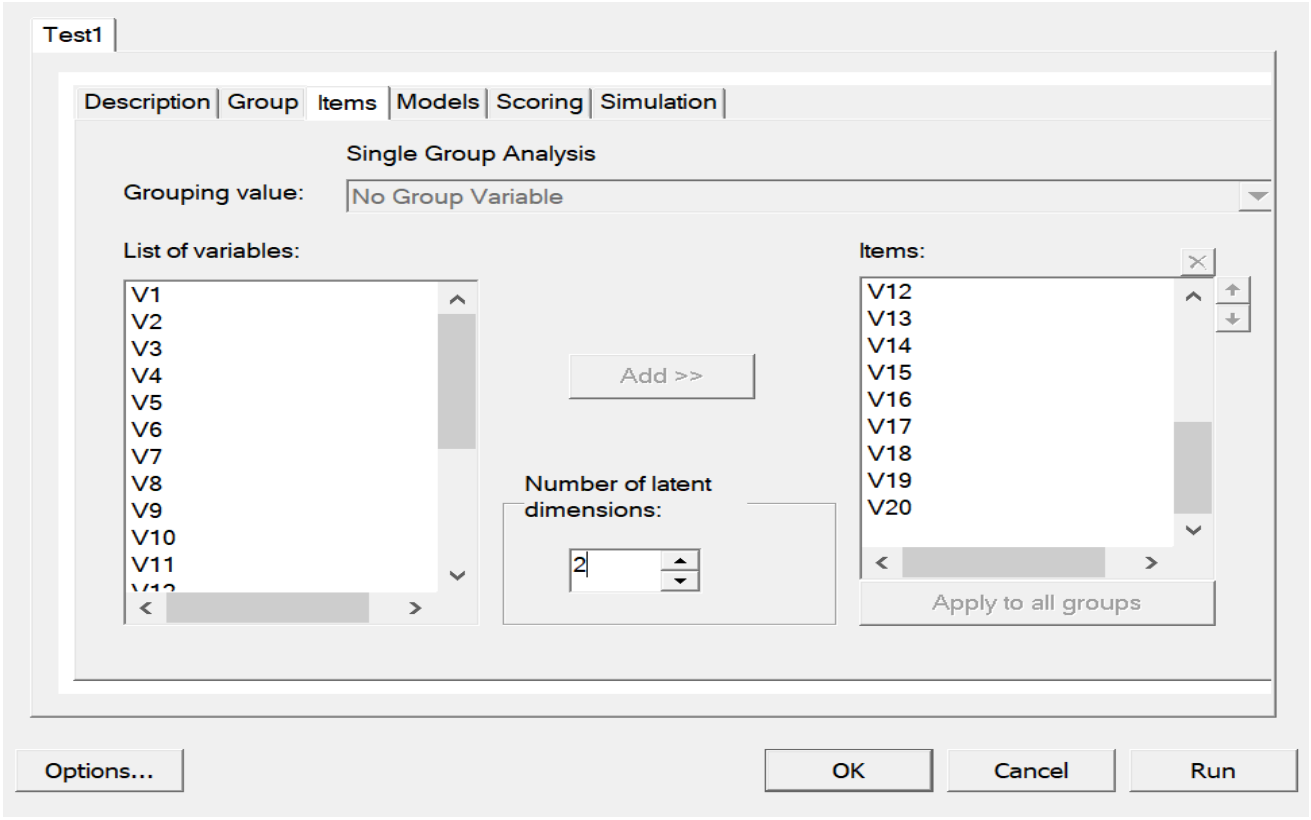
Step 1: Read data using .csv file to IRTPRO program. IRTPRO also supports other files (e.g., SPSS data file), but comma delimited files are the most suitable file type. The screenshot below shows the appearance after the data was successfully input.



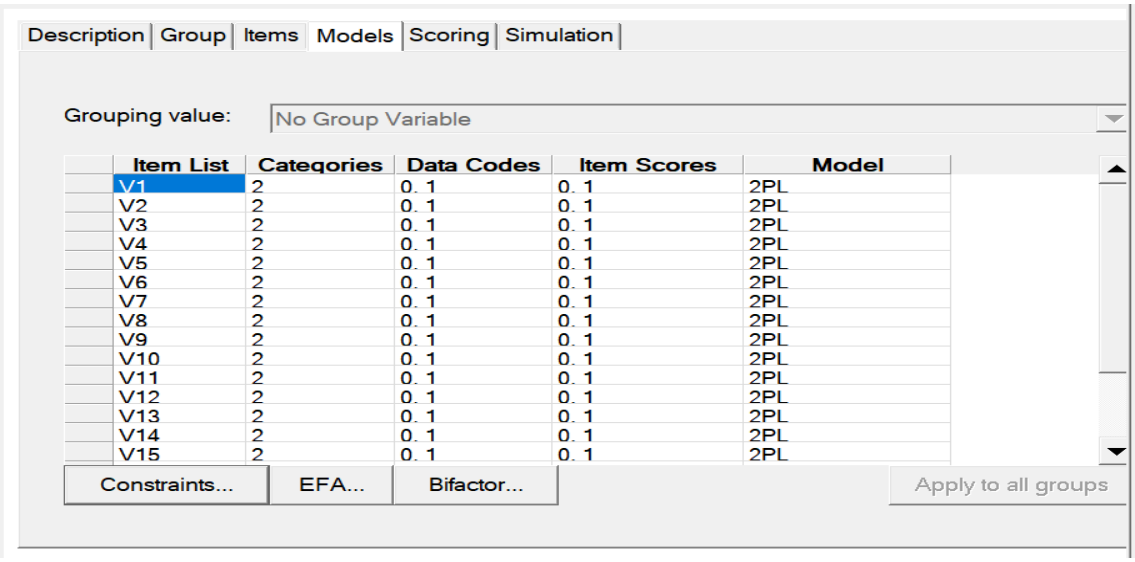
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
1	0	0	0	0	0	0	0	0	1	0	0
2	0	1	0	1	1	1	1	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	1	1	0	0	1	1	0	1
8	0	1	0	0	0	0	0	0	0	0	1
9	1	1	0	0	1	0	0	0	0	1	0

Step 2: Go to Model---Multidimensional IRT,

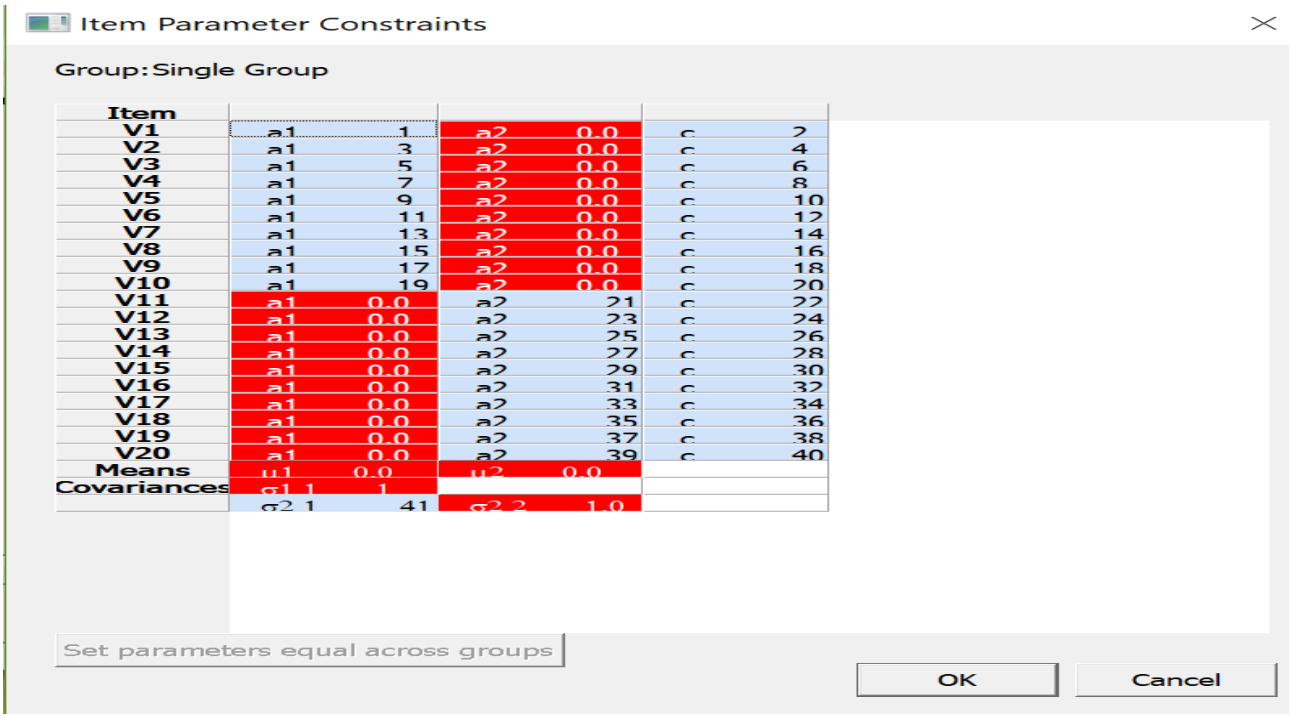
Click “item”, to specify the number of latent dimensions and which items will be calibrated. In this study, all 20 items will be moved to the right side by clicking “Add”



Step 3: click Model--Constraints,



Step 4: The screenshot below shows how the model should be constrained. Each cell is a parameter. The red color means "fixed"; the value of each parameter is either 0 or 1. The blue color means freely estimated. One can manually change this by right clicking the cell, then clicking OK.



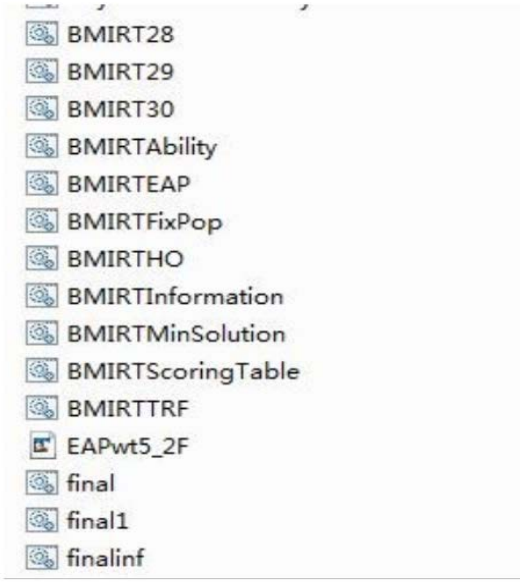
BMIRT

Step 1: On the "control file", input detailed information including the number of items, the number of examinees, and the number of dimensions. Please note: the order must exactly follow the order in the BMIRT Manual. The screenshot below shows the control file for this study.

```
500 20 1 1.0 1.0 10000 4000 2 2 923879631 0.0 1.0 0 1 0.00 1.00 1.50 0.01 0.0 1.00 0.01 100 400 0.01
1.0 0.0 0.01 1.0 0.0 0.05 0.05 0.1
22222222222222222222
1 10
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
11111111110000000000
00000000001111111111
1 0.1 0.1 1  0.0 0.0
```

Step 2: Read the data file. The file -- *rwo* file is the only data file that can be recognized by BMIRT. Also, please note: IRTPRO fixed the mean and variance at the dimensions, BMIRT fixed the first discrimination parameter for each dimension. Therefore, the mean and variance of the dimension are freely estimated.

Step 3: In BMIRT program, you can choose different programs in order to run different models. In this study, the model estimation requires choosing *final1.bat* by double-clicking on it.



Step 4: Find results that are stored in different output files: Item parameters are shown on "sim_1.par file". The correlation between both dimensions is shown on "sim_1 posterior.var file". Please make sure every file mentioned above is in the same folder, called "One-Group".