

**SOFTWARE PACKAGES FOR MULTIPLE GROUP IRT ANALYSIS AND  
ACCURACY OF PARAMETER ESTIMATES**

**Executive Report for the Maryland State Department of Education**

**By the MARC Team**

**(Bob Lissitz, Hong Jiao, Ming Li, Daniel Yangsup Lee and Yoonjeong Kang)**

**1. Multiple group IRT analysis**

- (1) Application of multiple group IRT analysis
  - Different groups with different test forms
    - Nonequivalent group horizontal equating of test forms
    - Vertical equating of test forms
  - Different groups with the same test form
    - DIF analysis & DRIFT analysis
    - Concurrent calibration of item parameters
    - Estimation and comparison of latent ability across groups
- (2) Main issues for multiple group IRT analysis
  - Model identification
    - Given that scale for latent ability in any IRT model is arbitrary, one must assign the metric of latent ability during calibration.
    - In a multiple group IRT analysis, many programs assign the metric of ability distribution by setting the mean and standard deviation of the ability distribution for one group (i.e., reference group) at mean of 0 and standard deviation of 1 while the mean and standard deviation of the ability distribution for the other groups (i.e., focal group) are freely estimated.
    - In addition, item parameters need to be constrained for model identification purposes. In general, for 1PL model at least one difficulty parameter across groups needs to be constrained while for 2 PL model, at least both one discrimination- and one difficulty parameter need to be constrained across groups.
  - Common scale or metric across different groups (Commonality)
    - Another issue that arises in multiple group IRT analysis is commonality. Commonality can be achieved when item and ability parameter estimates are on the same scale or metric across different groups.
    - Given that different groups have different ability levels, item parameters or ability parameters from separate IRT analysis can be on different scales. If item and ability parameter estimates are not on the common scale, comparisons across groups may not be meaningful.

- In general, a set of anchor items is selected and used to achieve commonality across different groups during calibration.

- (3) Available software programs for multiple group IRT analysis
- Commercial: BILOG-MG, MULTILOG, IRTPRO, flexMIRT, and Mplus
  - Noncommercial: BMIRT and FLIRT (R package).

## **2. Review of IRT software programs for multiple group IRT analysis**

### (1) Purpose of review

- Given that different software programs employ different defaults and/or options for model identification and commonality, it is important to understand what kinds of defaults and/or options each software program adopts for resolving the two main issues.
- The review of IRT software programs provides information particularly regarding these two main issues (model identification and commonality).
- The review focuses on use of software programs for multiple group IRT analysis in the context where the same test form is administered to different groups.
- Lastly, this review focuses on seven software programs for multiple group IRT analysis which include BILOG-MG, MULTILOG, IRTPRO, flexMIRT, Mplus, BMIRT, and FLIRT (R package).

### (2) Review summary

- Constraint on ability distribution for model identification purpose
  - BILOG-MG, IRTPRO, flexMIRT, and BMIRT fix the ability distribution of reference group (i.e., first group) to mean of 0 and standard deviation of 1 while mean and standard deviation of ability distribution for second group are freely estimated by default.
  - In MULTILOG, reference group is the second group and thus the ability distribution of the second group is fixed to mean of 0 and standard deviation of 1 by default. For the first group, standard deviation of the ability distribution is fixed to 1 while mean of ability distribution is freely estimated. The MULTILOG does not allow the switch of the first and second group. Users may do so manually by reversing the data for the two groups, such that the second group becomes the first group.
  - Mplus fixes first factor loading (i.e., discrimination parameter) to 1 across groups by default while means and standard deviation of the ability distribution for all groups are freely estimated.
  - FLIRT fixes the mean of the ability distribution for first group to 1 while mean of second group and standard deviations of both groups are freely estimated.
  - The defaults for model identification can be overridden easily by users with IRTPRO, flexMIRT, Mplus, BMIRT within the programs while the other programs are not flexible.

- Constraint on item parameter for model identification and commonality purposes
  - For multiple group IRT analysis, BILOG-MG, MULTILOG, and FLIRT automatically assume that all item parameters are the same for different groups and thus all items are treated as anchor items by default. Thus these programs provide only one set of item parameter by default.
  - With IRTPRO, flexMIRT, Mplus, and BMIRT, users can select anchor items and use them to achieve commonality. That is, these programs do not assume that all item parameters are the same for different groups.
  - With BILOG-MG and MULTILOG, users may trick the program to obtain group specific parameter estimates when employing multiple group IRT analysis. Users can do so by treating the same non-anchor items as different items across groups and change the file format accordingly.
  
- Model availability & data format
  - BILOG-MG and MULTILOG can estimate multiple group unidimensional IRT models only while the other software programs can estimate multiple group IRT model for both unidimensional and multidimensional models. Specific IRT models that each software program can estimate are presented in Appendix I.
  - Different software programs use different data formats. Particularly when BILOG-MG or MULTILOG is used for group-specific calibration, the non-anchor items should be treated as different test items and construct data file accordingly. In addition, flexMIRT requires separate group data files while the other programs require one data file for multiple group IRT analysis. More detailed information can be found in Appendix I.
  
- (3) Conclusions- Choice of software programs for multiple group IRT analysis
  - When the same test form is administered to different groups (i.e., different states), equality of all item parameters may not be tenable. In this case, use of multiple group IRT model is one approach to take group heterogeneity into account.
  - Seven software programs (BILOG-MG, MULTILOG, IRTPRO, MPLUS, flexMIRT, FLIRT (R package), BMIRT) are reviewed in this document and we found that IRTPRO, MPLUS, and flexMIRT seem to be more flexible software programs for multiple group analyses in terms of model identification and commonality. With these three programs, users can easily put item parameters on a common scale across groups.

### **3. Multiple group Rasch model analysis with simulated data**

- (1) Purpose of analysis
  - The purpose of the analysis was to (1) examine the impact of different choices of IRT models (single vs. multiple group) on item parameters and ability parameters and (2) evaluate available multiple group IRT software programs for multiple group IRT analysis.

## (2) Data simulation

- For the full scaled simulation study, the IRT models used in PARCC assessment will be selected.
- Manipulated factors in full scaled simulation study have 9 factors:
  - Number of groups (2 groups; reference group vs. focal group)
  - Sample size (3 levels: 1,000, 2,000, and 4,000 per group)
  - Number of total items (50 items)
  - Number of anchor items (15 items)
  - Number of DIF items (2 levels: 5 items (10%) and 10 items (20%))
  - Magnitude of DIF (2 levels: 0.5 and 1.0)
  - Type of DIF (2 levels: uniform vs. non-uniform)
  - Mean ability difference between two groups (0.5; Group1=0 vs. Group2= - 0.5)
  - Standard deviation difference in ability distribution across two groups (2 levels: 0 and 0.5).
- As a preliminary simulation, one data set following the RASCH model was simulated under the following conditions and analyzed:
  - Number of groups (2 groups)
  - Sample size (4,000 per group)
  - Number of total items (50 items)
  - Number of anchor items (15 items)
  - Number of DIF items (5 items=10%)
  - Magnitude of DIF (0.5)
  - Type of DIF (uniform)
  - Mean ability difference between two groups (0.5)
  - Standard deviation difference in ability distribution across two groups (0=no difference).

## (3) Preliminary Data Analysis

- The software program *R* (R Development Core Team, 2010) was used to generate item response data following the Rasch model. For the reference group, the true item difficulty parameters and ability distribution were generated from a normal distribution with mean of 0 and standard deviation of 1. So this is the matched case between ability distribution and item difficulty parameters. The DIF item parameters for the focal group were simulated by subtracting 0.5 from each reference group item's difficulty parameter. Generated item difficulty parameters for both groups are presented in Table 1 in Appendix II.
- With simulated data, three different Rasch analyses were performed (single group, multiple group with equality constraint on all items, vs. multiple group analysis with equality constraint on anchor items) using seven different software programs.
- For model identification, the ability distribution for the reference group was set at mean of 0 and standard deviation of 1 while the ability distribution for the focal group was freely estimated.
- To obtain group specific item parameters with BILOG-MG and MULTILOG, we changed the data file such that the same non-anchor items treated as different items across groups.

- For the MULTILOG program, we manually reversed the data for two groups in data file so that reference group matched other software programs.

#### (4) Results

- When ignoring group heterogeneity in item parameters and using a single group IRT analysis for calibration, the results revealed that the single group IRT analysis produced large amount of biases of both item- and ability parameter estimates.
- When comparing two multiple group IRT analyses, the results provided similar results. This result might be because both the number of DIF item and the magnitude of DIF were relatively small.
- From the software program perspective, BILOG-MG, IRTPRO, Mplus, and flexMIRT provided similar results across three analyses in terms of item- and ability parameter estimates. MULTILOG, FLIRT, and BMIRT provided dissimilar results from BILOG-MG, IRTPRO, Mplus, and flexMIRT. The factors affecting these results will be examined in full scaled simulation study.
- The bias values of the item and ability parameter estimates are presented in Table 2 and 3 in Appendix II.

#### (5) Implication

- When the same test form is administered to different groups (e.g., states) like assessment for Common Core State Standards, it is typically the case that group homogeneity cannot hold. Therefore, it is important to take such group heterogeneity into account during the process of calibration and ability estimation.
- The preliminary results support that ignoring group heterogeneity and using single group IRT analysis may produce biases in the item and ability parameter estimates and thus may not provide fair results and accurate measures of students' achievement. The results support that multiple group IRT analysis produced lower biases in item and ability parameter estimates.
- Multiple group IRT model could be one useful approach for analyzing item response data from different populations by taking group heterogeneity into account.
- In the full scaled simulation study, the IRT models used in PARCC assessment will be chosen for the simulation. Rigorous examination will be conducted to (1) investigate the consequences of ignoring group heterogeneity in item and ability parameter estimates under various conditions, and (2) evaluate available multiple group IRT software programs with a large number of replicated data set.

## Appendix I.

### Software Programs for Multiple Group IRT Analysis

<b>BILOG-MG</b> (Michele Zimowski, Eiji Muraki, Robert Mislevy & Darrell Bock)			
Multi-Group IRT Model	Model Identification	Commonality	Note
<p>Fits</p> <ul style="list-style-type: none"> <li>▪ Multiple group 1PL, 2PL, and 3PL models to binary data</li> </ul>	<ul style="list-style-type: none"> <li>▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of 1) while mean and standard deviation of the ability distribution for second group are freely estimated.</li> <li>▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.</li> </ul>	<ul style="list-style-type: none"> <li>▪ By putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric.</li> <li>▪ In case where users want to use a set of anchor items to achieve commonality, user should use the trick. User can do so by changing data file format. That is, users should treat the same non-anchor items as different items across groups and change the file format accordingly.</li> </ul>	<ul style="list-style-type: none"> <li>▪ BILOG-MG provides Graphical User Interface(GUI).</li> <li>▪ BILOG-MG can handles omits and not-presented data</li> <li>▪ For DIF analysis, BILOG-MG estimate item parameters separately for each group. Then the item parameters are adjusted under the constraint that the mean thresholds of the groups to be equal. However, BILOG-MG implements DIF/DRIFT analyses for only 1PL model.</li> <li>▪ Obtaining group-specific item parameter estimates is not easy because users need to change data format outside of the BILOG-MG program.</li> </ul>
<b>MULTILOG</b> (David Thissen)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
<p>Fits</p> <ul style="list-style-type: none"> <li>▪ Multiple-group 1PL, 2PL, or 3PL to binary data</li> <li>▪ Multiple-group graded response model, nominal model, and multiple choice model to polytomous data</li> </ul>	<ul style="list-style-type: none"> <li>▪ It automatically fixes the mean and standard deviation of ability distribution for second group (mean of 0 and standard deviation of 1) while for first group, standard deviation of the ability distribution is fixed to one and mean of the ability distribution is freely estimated.</li> <li>▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Similar to BILOG-MG, by putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric.</li> <li>▪ In case where users want to use a set of anchor items to achieve commonality, user should use the trick. User can do so by changing data file format. That is, users should treat the same non-anchor items as different items across groups and change the file format accordingly.</li> </ul>	<ul style="list-style-type: none"> <li>▪ It provides GUI.</li> <li>▪ With common item method, it needs to be run twice for DIF analysis.</li> <li>▪ Like BILOG-MG, obtaining group-specific item parameter estimates is not easy because users need to change data format outside of the MULTILOG program.</li> <li>▪ It provides DIF analysis.</li> </ul>
<b>IRTPRO</b> (Cai, Thissen & du Toit, 2011)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
<p>Fits</p> <ul style="list-style-type: none"> <li>▪ Multi-group 1PL, 2PL, or 3PL logistic model to binary data</li> <li>▪ Multi-group rating</li> </ul>	<ul style="list-style-type: none"> <li>▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of one) while mean</li> </ul>	<ul style="list-style-type: none"> <li>▪ There is no default for commonality.</li> <li>▪ It is easy for users to select the anchor items and constrain parameters of the anchor items</li> </ul>	<ul style="list-style-type: none"> <li>▪ It provides GUI.</li> <li>▪ It provides DIF or drift analysis for various IRT models including unidimensional and multidimensional model.</li> </ul>

scale model, partial credit model, generalized partial credit, graded response model, or nominal model to polytomous data	and standard deviation of the ability distribution for second group are freely estimated. ▪ It is easy for users to change default option for the model identification. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	to be equal across groups for achieving commonality.	▪ Generally, it is easy to run IRT models.
<b>flexMIRT</b> , Version 2.0 (Cai,2013)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple-group 1PL, 2PL, and 3PL logistic model ▪ multi-group generalized partial credit model, partial credit model, and rating scale model ▪ multi-group bifactor model ▪ multilevel and multidimensional models of all the preceding models	▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of 1) while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ It is easy for users to change default option for the model identification. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.	▪ It provides DIF or drift analysis for various IRT models including unidimensional and multidimensional model. ▪ For multiple group IRT analysis, data for each group should be stored as individual files.
<b>Mplus</b> (Muthén, & Muthén, 1998-2014)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
Fits ▪ multiple group 1PL and 2PL logistic model to binary data ▪ multi-group Samejima's graded response model to polytomous data ▪ multilevel and multidimensional models of all the preceding models	▪ It automatically fixes the first factor loading (i.e., discrimination parameter) to one across groups by default while means and standard deviations of the ability distributions for all groups are freely estimated. ▪ It is easy for users to change default option for the model identification. That is, users can fix the mean and standard deviation of the ability distribution for first group while mean and standard deviation of the ability distribution for second group are freely estimated. ▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.	▪ There is no default for commonality. ▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.	▪ It uses different parameterization from other IRT software programs. For difficulty parameters, users can obtain by dividing threshold parameters in Mplus output by corresponding factor loading. ▪ It provides DIF or drift analysis for 1PL, 2PL, and graded response model. ▪ 3PL model cannot be estimated with Mplus.

<b>BMIRT</b> (Lihua Yao, 2003,2010)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
<p>Fits</p> <ul style="list-style-type: none"> <li>▪ multiple-group 1PL, 2PL, or 3PL logistic model to binary data</li> <li>▪ multiple-group generalized two-parameter partial credit model, the testlet model, the graded response model, and the higher-order IRT model.</li> <li>▪ multidimensional model of the above models</li> </ul>	<ul style="list-style-type: none"> <li>▪ It automatically fixes the mean and standard deviation of the ability distribution for first group (mean of 0 and standard deviation of 1) while mean and standard deviation of the ability distribution for second group are freely estimated.</li> <li>▪ Users have an option for choosing a set of anchor items and putting equality constraints on anchor items across groups.</li> </ul>	<ul style="list-style-type: none"> <li>▪ There is no default for commonality.</li> <li>▪ It is easy for users to select the anchor items and constrain parameters of the anchor items to be equal across groups for achieving commonality.</li> </ul>	<ul style="list-style-type: none"> <li>▪ It is non-commercial program. Users need to email to author and then download them program.</li> <li>▪ It uses Bayesian estimation. Users need to specify information on parameters of the priors and proposals in code. So users who don't know Bayesian estimation, it would not be easy to use BMIRT software.</li> <li>▪ For IRT analysis, users need to install and specify the path JAVA environmental variable on computer <ul style="list-style-type: none"> <li>▪ The program code is quite different from other IRT program.</li> </ul> </li> </ul>
<b>FLIRT</b> (Jeon, Rijman, & Rabe-Hesketh, 2014)			
Multiple-Group IRT Model	Model Identification	Commonality	Note
<p>Fits</p> <ul style="list-style-type: none"> <li>▪ multiple-group 1PL, 2PL, and bifactor model</li> <li>▪ multidimensional models of all the preceding models</li> </ul>	<ul style="list-style-type: none"> <li>▪ It fixes the mean of the ability distribution for first group (mean of 0) while mean of the ability distribution for second group is freely estimated. Standard deviations of the ability distributions for all groups are estimated.</li> <li>▪ It constrains all item parameters to be equal for different groups by default when employing multiple group IRT analysis.</li> </ul>	<ul style="list-style-type: none"> <li>▪ By putting equality constraints on all item parameters, it assumes that item and ability parameter estimates are on the same scale or metric.</li> <li>▪ According to author, she will going to add an option for putting equality constraints on anchor items soon.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Item covariates or person covariates can be included in IRT model.</li> <li>▪ It provides DIF analysis.</li> <li>▪ With the current version, FLIRT cannot be used for 3PL and polytomous IRT models.</li> <li>▪ Installation process is complicated. "FLIRT" package requires the Matlab Compiler Runtime (MCR) and it should be downloaded and installed before installing "FLIRT" package.</li> </ul>



**Appendix II.**  
Simulation Study of Multiple Group IRT Analysis

Table 1  
*True Difficulty Parameter in Reference and Focal Group*

Item #	Difficulty Parameter (Reference Group)	Difficulty Parameter (Focal Group)	Item #	Difficulty Parameter (Reference Group)	Difficulty Parameter (Focal Group)
1	-0.136	-0.136	26	<b>0.234*</b>	<b>0.734*</b>
2	-0.041	-0.041	27	<b>0.593*</b>	<b>1.093*</b>
3	1.011	1.011	28	<b>2.001*</b>	<b>2.501*</b>
4	-0.158	-0.158	29	<b>-1.837*</b>	<b>-1.337*</b>
5	-2.157	-2.157	30	<b>-0.862*</b>	<b>-0.362*</b>
6	0.499	0.499	31	1.583	1.583
7	-0.755	-0.755	32	0.155	0.155
8	0.779	0.779	33	-0.275	-0.275
9	0.755	0.755	34	0.788	0.788
10	-1.100	-1.100	35	-0.223	-0.223
11	0.167	0.167	36	1.392	1.392
12	-0.029	-0.029	37	-0.489	-0.489
13	1.876	1.876	38	0.137	0.137
14	0.245	0.245	39	0.004	0.004
15	0.702	0.702	40	-0.727	-0.727
16	-0.015	-0.015	41	-0.721	-0.721
17	-0.143	-0.143	42	-0.191	-0.191
18	0.321	0.321	43	1.335	1.335
19	0.122	0.122	44	0.356	0.356
20	-0.595	-0.595	45	0.843	0.843
21	-0.442	-0.442	46	0.775	0.775
22	0.291	0.291	47	0.080	0.080
23	0.724	0.724	48	-0.673	-0.673
24	0.460	0.460	49	1.836	1.836
25	0.185	0.185	50	-0.207	-0.207

*Note.*

\* indicate DIF items.

Table 2  
*Comparison of Bias of Difficulty Parameter Estimates across Three Analyses*

Software	Single Group Analysis <sup>1</sup>	Multiple Group Analysis- Equality constraint on all items <sup>1</sup>	Multiple Group Analysis- Equality constraint on anchor items
Average of bias across all items (SD)			
BILOG-MG	0.154 (0.163)	-0.042 (0.151)	R: 0.008 (0.033) F: 0.031 (0.039)
MULTILOG	<0.001 (7.223)	<0.001 (0.080)	R: <0.001 (0.067) F: -0.091 (0.159)
IRTPRO	0.281 (0.078)	0.007 (0.078)	R: 0.005 (0.030) F: 0.021 (0.038)
Mplus	0.262(0.086)	0.002 (0.078)	R: 0.001(0.034) F: 0.015 (0.044)
flexMIRT	0.280 (0.078)	0.006 (0.078)	R: 0.003 (0.030) F: 0.019 (0.038)
FLIRT	-0.617 (1.692)	-0.338 (1.693)	-
BMIRT	0.073 (0.072)	-0.072 (0.077)	R: 0.006(0.006) F: 0.131(0.053)

*Note.*

True item parameters for focal group were used to calculate bias in single group analysis and multiple group analysis with equality constraint on all items.

Table 3  
*Comparison of Bias of Ability Estimates across Three Analyses*

Software	Single Group Analysis <sup>1</sup>	Multiple Group Analysis- Equality constraint on all items	Multiple Group Analysis- Equality constraint on anchor items
Average of bias across all ability (SD)			
BILOG-MG	R: 0.177(0.325) F: 0.332 (0.524)	R: 0.002(0.325) F: 0.156(0.524)	R: 0.002 (0.325) F: 0.074 (0.336)
MULTILOG	R: -0.531(0.346) F: -0.501 (0.352)	R: -0.002 (0.317) F: -0.038 (0.324)	R: -0.041 (0.317) F: -0.042 (0.324)
IRTPRO	R: 0.248(0.313) F: 0.261 (0.324)	R: 0.001(0.313) F: -0.043(0.325)	R: 0.001 (0.313) F: 0.017 (0.325)
Mplus	R: 0.243(0.318) F: 0.278 (0.328)	R: 0.004(0.316) F: -0.026(0.326)	R: 0.003 (0.316) F: 0.035(0.326)
flexMIRT	R: 0.247(0.313) F: 0.260 (0.324)	R: <0.001(0.313) F: -0.045(0.325)	R: <0.001 (0.313) F: 0.015 (0.325)
FLIRT	R: 0.238 (0.317) F: 0.263 (0.326)	R: 0.007 (0.316) F: 0.512 (0.325)	-
BMIRT	R: 0.050(0.326) F: 0.066 (0.336)	R: -0.073 (0.326) F: -0.120 (0.338)	R: -0.008 (0.329) F: 0.161 (0.341)

## References

- Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.
- Jeon, M., Rijmen, F. and Rabe-Hesketh, S. (2014). flirt: Flexible Item Response Theory Modeling. R package version 1.13. <http://www.mathworks.com/products/compiler/mcr/>
- Muthén, B., & Muthén, L. K. (1998-2014). MPLUS (Version 7.11). [Computer Software]. Los Angeles, CA: Muthén & Muthén.
- Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software International
- Yao, L. (2010). BMIRTII: Bayesian multivariate item response theory—second version. [Computer software]. Monterey, CA: *www.BMIRT.com*.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.