Comparative Analyses of MIRT Models and Software (BMIRT and flexMIRT)

Educational and Psychological
Measurement

2017, Vol. 77(2) 263–274

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416661220

journals.sagepub.com/home/epm

Guler Yavuz¹ and Ronald K. Hambleton²

Abstract

Application of MIRT modeling procedures is dependent on the quality of parameter estimates provided by the estimation software and techniques used. This study investigated model parameter recovery of two popular MIRT packages, BMIRT and flexMIRT, under some common measurement conditions. These packages were specifically selected to investigate the model parameter recovery of three item parameter estimation techniques, namely, Bock–Aitkin EM (BA-EM), Markov chain Monte Carlo (MCMC), and Metropolis–Hastings Robbins–Monro (MH-RM) algorithms. The results demonstrated that all estimation techniques had similar root mean square error values when larger sample size and higher test length were used. Depending on the number of dimensions, sample size, and test length, each estimation technique exhibited some strengths and weaknesses. Overall, the BA-EM technique was found to have shorter estimation time with all test specifications.

Keywords

multidimensional item response theory, BMIRT, flexMIRT, item parameter recovery, MCMC, MH-RM, BA-EM

MIRT Parameter Estimation Software and Techniques

Demand for applying multidimensional item response theory (MIRT) models to educational test data has risen in recent years. Applications are being considered by major test publishers and other testing agencies to improve the analysis of

¹Adiyaman University, Adiyaman, Turkey

²University of Massachusetts Amherst, MA, USA

Corresponding Author:

Guler Yavuz, Department of Educational Sciences, Adiyaman University, Adiyaman, Turkey.

Email: gyavuz2010@gmail.com

achievement test data associated with educational reform and ability testing. However, serious questions remain about modeling, parameter estimation, and the available software itself. The application of MIRT model procedures is dependent on the quality of model parameter estimates that are obtained from the estimation techniques used in the software. Several questions have been raised about the quality of these estimates obtained from some of the MIRT software. The new software programs that are available vary in the estimation procedures used and the ways that they address indeterminacy issues.

Some researchers have designed and developed new MIRT software packages (such as BMIRT, MIRT, IRTPRO, and flexMIRT) that provide different estimation techniques. For example, BMIRT (Yao, 2003) (Bayesian multivariate item response theory) that implements Markov chain Monte Carlo (MCMC) methods applies the Metropolis–Hastings sampling algorithm to estimate the item, examinee, and population distribution parameters for a set of MIRT models for both dichotomously and polytomously scored test items. flexMIRT implements BA-EM and MH-RM algorithms to estimate the item, examinee, and population distribution parameters for data sets that have different test conditions. In addition to these, IRTPRO 2.1 software program (Cai, du Toit, & Thissen, 2011) has features very similar to flexMIRT. Therefore, different options are available when a researcher intends to decide on software for analyzing data set.

The aim of this study is to investigate and compare model recovery performance of different packages that used different estimation techniques. flexMIRT and BMIRT were used in this study. These software packages have been preferred because of their features. For example, BMIRT, which is free, is one of the popular software packages that include MCMC estimation technique. flexMIRT is the other popular software that can be used for parameter estimation of MIRT models. Two different estimation techniques can be used with flexMIRT: BA-EM and MH-RM. According to several model parameter recovery studies using IRT and MIRT, the recovery of parameters that have been estimated by different software produce different root mean square error (RMSE) results with different test specifications (see, Bahry, 2012; Beguin & Glas, 2001; Bolt & Lall, 2003; Cai, 2010a; Cai & Monroe, 2013; Chalmers, 2012; De la Torre, 2009; Finch, 2011; Finch & French, 2012; Guyer & Thompson, 2011; Hsieh, Proctor, Hou, & Teo, 2010; Jurich & Goodman, 2009; Kieftenbeld & Natesan, 2012; J. Lee, 2012; S. H. Lee, 2007; Montgomery & Skorupski, 2012; Svetina, 2011; Weiss & Minden, 2012; Yao & Schwarz, 2006). Because flexMIRT, IRTPRO, BMIRT, and some MIRT packages in R are more popular, this study has been completed in two stages. First of all, the MIRT parameters were estimated with flexMIRT, IRTPRO, BMIRT, and R. According to the results, similar parameter values were estimated with flexMIRT and IRTPRO. Because flexMIRT has new features, it was preferred in the study. R packages were not used because of some problems with test specifications and computing speed. The second package that has been used in this study is BMIRT, which is one of the best option that MCMC can be applied with.

The major challenge in parameter estimation techniques in MIRT models is that the relationship between parameter recovery and test specifications such as the number of dimensions, dimension structure, number of items, number of examinees, and selection of parameter distributions is still unclear.

This study is constructed to provide multidimensional item parameter recovery with two multidimensional models, two software programs, and three estimation techniques. These are, respectively, the compensatory multidimensional three-parameter model and the multidimensional graded response model, BMIRT (MCMC with Metropolis–Hasting method), flexMIRT (with Bock–Aitkin EM with generalized dimensional reduction [BA-EM]; Bock & Aitkin, 1981), and Metropolis–Hastings Robbins–Monro (MH-RM; Cai, 2010b, 2010c) algorithm. The research questions addressed in this study include the following: Does (1) the number of items, (2) the sample size, and (3) the number of dimensions affect item parameter recovery when the MC3PL and MGRM models are calibrated with MCMC, BA-EM, and MH-RM; and how does (4) estimation time compare with these various test specifications?

Method

Data Generation

Ability parameters were generated from a multivariate normal distribution in which each ability dimension had a mean of 0 and a variance of 1. The ability parameters were drawn from a multivariate normal distribution with a prespecified variance–covariance matrix, $\theta_i \sim MVN(0, \Sigma)$. For the three-dimensional model and the five-dimensional model, the mean and variance vectors were $\mu = \{0, 0, 0\}$ $\sigma = \{1, 1, 1\}$; $\mu = \{0, 0, 0, 0, 0\}$ $\sigma = \{1, 1, 1, 1, 1\}$.

Item parameters of MC3PL item responses and multidimensional graded responses were generated by WinGen (Han, 2007) and R. To generate the parameters, different distributions were used. a (discrimination) and d (difficulty) parameters were generated from a uniform distribution, and the c (guessing) parameter was generated from a beta distribution. These values are provided in Table 1.

Simulation Conditions

In this study, before deciding on the simulation conditions, first the result of previous simulation studies were investigated (see, Bolt & Lall, 2003; Bulut, 2013; De la Torre, Song, & Hong, 2011; J. Lee, 2012; Yao & Boughton, 2007; B. Zhang & Stone, 2004, 2007; J. Zhang, 2012), and then, the item parameters of big test applications (such as MCAS) were examined. Following that, the distribution of parameters and sample sizes, test lengths, and dimensions were selected. After generating item parameters, both dichotomous and polytomous item responses were simulated with different test conditions: test length (15, 30, 60), sample size (1,000, 2,000, 5,000), the number of dimensions (3, 5), and the correlation between dimensions (0.00, 0.30,

Table 1. Distribution Values and Types of Parameters.

Test structure	Models	<i>a</i> (Uniform)	<i>d</i> (Uniform)	<i>c</i> (Beta)
Simple structure	Multidimensional three parameter model (MC3PL)	(0.6)-(1.4)	(-2)-(2)	(100)-(400)
	Multidimensional graded response model (MGRM)	(0.4)-(1.2)	(-2)-(2)	—
Bifactor model	MC3PL	Primary dimension	(0.6)-(1.4)	(-2)-(2)
		Specific dimensions	(0.4)-(1.2)	(-2)-(2)
	MGRM	Primary dimension	(0.4)-(1.2)	(100)-(400)
		Specific dimensions	(0.2)-(1.0)	(-2)-(2)

0.60, 0.90) in SimuMIRT (Yao, 2003). Some likely test data sets in practice were created mimicking three and five dimensions, three test lengths, and three sample sizes. Data sets were simulated with different types of test structure (simple structure model and bifactor model structure).

Twenty-one (equally spaced between -5 and 5) quadrature points per dimension were chosen for the BA-EM method. For the intercept, a lognormal distribution (1.0, 0.5) and a beta prior distribution (100, 400) were chosen. The default values were used for MH-RM techniques but processors, imputations, proposal std, and initial gain values were used as 4.0, 2.0, 0.2, and 0.1, respectively. For MCMC, the same priors that were used with BA-EM were used. (BMIRT values, for one analyse, sample values are such as: 8,000, 1,000, 5, 2, 923879631, 0.0, 1.0, 0.2, 0.1, 0.03, 1.0, 0.5, 0.01, 0.0, 1.0, 0.01, 100, 400, 0.01, 1.0, 0.0, 0.01, 1.0, 0.0, 0.05, 0.05, and 0.1.) In this study, one of the important limitations is the values of priors that were used with BMIRT and flexMIRT. Because these values have an important effect on the estimation time, different values of priors were tested and then the values by which the optimum values of parameters estimated were preferred for both software.

The iteration number was preferred as 8,000 and burn in was preferred as 1,000 for MCMC. One of the limitations of this study is the iteration number that was used. Before deciding the number of iteration, two values of 4,000 and 8,000 were tested and because 8,000 value provided better results, this value was preferred. But according to our expectation, the higher values might provide better results and more comparable results. To evaluate the accuracy of item parameter recovery, RMSE was used. An RMSE was computed for each parameter and then averaged across parameters.

Finally, the estimation time for the two software packages under each condition was computed. Computation time was averaged over 25 replications.

Results

Table 2 and Figures 1 and 2 show the RMSE values of multidimensional compensatory three-parameter logistic model that has simple structure. In Table 2, the RMSE

Table 2. RMSE Values for the Multidimensional Compensatory Three-Parameter Model (Simple Structure).

		Three dimensions									Five dimensions								
		1,000			2,000			5,000			1,000			2,000			5,000		
		15	30	60	15	30	60	15	30	60	15	30	60	15	30	60	15	30	60
MCMC	a1	0.26	0.18	0.16	0.15	0.14	0.11	0.10	0.09	0.07	0.31	0.13	0.10	0.27	0.18	0.13	0.21	0.11	0.09
	a2	0.31	0.26	0.17	0.22	0.20	0.16	0.16	0.14	0.12	0.35	0.15	0.10	0.30	0.17	0.13	0.18	0.10	0.08
	a3	0.28	0.17	0.15	0.18	0.14	0.11	0.12	0.09	0.08	0.33	0.15	0.12	0.29	0.16	0.12	0.22	0.11	0.08
	a4										0.31	0.20	0.13	0.32	0.15	0.12	0.24	0.11	0.08
	a5										0.33	0.24	0.19	0.29	0.17	0.14	0.28	0.13	0.09
MH-RM	d	0.14	0.12	0.10	0.11	0.10	0.10	0.09	0.08	0.08	0.13	0.11	0.10	0.12	0.10	0.09	0.10	0.08	0.07
	c	0.03	0.03	0.01	0.03	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	a1	0.22	0.17	0.16	0.15	0.14	0.12	0.10	0.09	0.08	0.29	0.20	0.19	0.24	0.17	0.14	0.14	0.11	0.10
	a2	0.28	0.27	0.15	0.18	0.14	0.12	0.11	0.10	0.10	0.32	0.22	0.18	0.23	0.16	0.14	0.14	0.10	0.08
	a3	0.24	0.18	0.17	0.17	0.14	0.13	0.12	0.10	0.09	0.25	0.21	0.18	0.22	0.15	0.14	0.17	0.10	0.08
BA-EM	a4										0.27	0.20	0.18	0.24	0.14	0.18	0.17	0.11	0.08
	a5										0.30	0.21	0.10	0.24	0.16	0.20	0.20	0.13	0.10
	d	0.14	0.12	0.10	0.11	0.10	0.09	0.09	0.09	0.08	0.13	0.12	0.11	0.11	0.10	0.09	0.09	0.08	0.07
	c	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	a1	0.23	0.17	0.16	0.15	0.14	0.11	0.11	0.10	0.09	—	—	—	—	—	—	—	—	—
	a2	0.28	0.27	0.16	0.18	0.16	0.10	0.14	0.12	0.10	—	—	—	—	—	—	—	—	—
	a3	0.24	0.18	0.17	0.17	0.14	0.12	0.12	0.11	0.10	—	—	—	—	—	—	—	—	—
	a4										—	—	—	—	—	—	—	—	—
	a5										—	—	—	—	—	—	—	—	—
	d	0.14	0.12	0.10	0.11	0.10	0.10	0.09	0.09	0.08	—	—	—	—	—	—	—	—	—
	c	0.04	0.03	0.01	0.03	0.01	0.01	0.01	0.01	0.01	—	—	—	—	—	—	—	—	—

Figure 1. RMSE values for the multidimensional compensatory three-parameter model (simple structure, three dimension).

Figure 2. RMSE values for the multidimensional compensatory three parameter model (simple structure, five dimension).

values of all the estimation techniques (MCMC, MH-RM, and BA-EM) were given with all the test conditions (three and five dimensions; 1,000, 2,000, and 5,000 sample size; 15, 30, and 60 items; the correlation between dimensions is 0.30).

Sample size especially had an important effect on all the estimation techniques. When the sample size was 5,000 and the number of items was 60, the RMSE values were the smallest. When sample size was increased from 1,000 to 2,000 and 2,000 to 5000, all the RMSE values decreased, as would be expected. For example, when the number of items was 15, 30, and 60, the RMSE values which were computed from *a1* parameters with the MCMC technique were given respectively; the RMSE values

of $a1$ parameters were computed as follows: for 15 items, they were 0.26, 0.15, 0.10; for 30 items, they were 0.18, 0.14, 0.09; and for 60 items, they were 0.16, 0.11, 0.07. With 15 items and a sample size of 1,000, the RMSE values which were computed from $a1$ parameters of MH-RM and BA-EM techniques were smaller than the MCMC technique. For d , and especially c parameters, the RMSE values of MCMC, MH-RM, and BA-EM were very close in all the test conditions.

In Table 2 and Figures 1 and 2, the RMSE values of the five-dimensional data sets were given for all the test conditions. The RMSE values of $a1$, $a2$, $a3$, d , and c parameters could not be computed for all the test conditions. Because the computer that was used for estimation had not enough computation capacity, the parameters of the five-dimensional data sets could not be estimated with the BA-EM algorithm. The RMSE values of MCMC and MH-RM techniques are given in Table 2 and Figure 2. According to the results, with 15 items and an examinee sample size of 1,000, a parameters ($a1$, $a2$, $a3$, $a4$, $a5$), which were estimated from five-dimensional data sets, had higher RMSE values than the a parameters ($a1$, $a2$, $a3$), which were estimated from the three-dimensional data sets. But higher sample sizes and item numbers provide the same RMSE values for both MCMC and MH-RM. The RMSE values of the d and c parameters, which were estimated from the five-dimensional data sets, were very close for both estimation techniques and data sets. Yet the RMSE values of parameters were very close to MH-RM and BA-EM.

In this study, the model parameter recovery of the multidimensional three-parameter logistic model and the multidimensional graded response model were investigated with two different model structures: simple and bifactor model structures. RMSE values of the multidimensional three-parameter logistic model are given in Table 2 and Figures 1 and 2. The RMSE values of the multidimensional three-parameter logistic model that has bifactor model structure and the RMSE values of the multidimensional graded response model that has simple structure (with different correlation values) and that has bifactor model structure were tested and computed.

Estimation Time of MCMC, MH-RM, and BA-EM

Figure 3 (the data set that has three dimensions) and Figure 4 (the data set that has five dimensions) show the average estimated time in seconds for the multidimensional compensatory three-parameter logistic model that has simple structure. Twenty-five replications were used for all the test conditions. According to all tables, the estimation time of BA-EM was smaller than those of MCMC and MH-RM. MCMC also had the highest estimation time for all of the conditions.

The estimation times of BA-EM were 19, 30, 67, 21, 60, 116, 90, 148, and 270 seconds, respectively; the estimation times of MH-RM were 74, 104, 140, 78, 263, 305, 422, 512, and 632 seconds, respectively, and the estimation times of MCMC were 96, 139, 240, 175, 281, 456, 488, 683, 1164 seconds, respectively, for all the test conditions that are given in Figure 3. Similar results have been gained through five-

Figure 3. Average estimation time (in seconds) for the multidimensional three-parameter model (simple structure, three dimension).

Figure 4. Average estimation time (in seconds) for the multidimensional three-parameter model (simple structure, five dimension).

dimensional data sets. The estimation time of MH-RM and MCMC are close to each other in some test specifications (when the data set has 15 items and 1,000 examinees), but according to the major part of the results, the estimation time of MCMC was higher than that of the MH-RM. The data sets with higher sample size and test had higher estimation time. For example, the average estimation time of MCMC with one of data sets that included 15 items and 1,000 examinees was 96 seconds. The

average estimation time of MCMC with another data set that included 60 items and 5,000 examinees was 1,164 seconds.

Discussion

There is substantial demand for multidimensional models and their application for building vertical scales, equating test scores, and subtest score reporting. However, serious questions remain about the modeling, parameter estimation, and the MIRT software itself. Based on the many findings from this study, BA-EM, MH-RM, and MCMC have advantages as well as shortcomings for data sets that have different properties. BA-EM is practical because of the shorter estimation time for applications, but if the researchers have not suitable computing power the item parameters for five-dimensional solutions cannot be estimated with. All the analyses were performed on a laptop with Intel Core i7-4790 with CPU@ 3.60 GHz processors with 8 GB of RAM. So, if researchers are interested in estimating the five-dimensional data sets' parameters with flexMIRT (BA-EM), they should choose more powerful computing equipment. MCMC is very suitable for complex modeling, which has broad sample sizes and test for estimation, but MCMC with BMIRT or other software take a lot of time, relatively speaking. These findings have been confirmed by others (see, Bolt & Lall, 2003; Bulut, 2013; B. Zhang & Stone 2004).

The use of MCMC and MH-RM in a MIRT framework is relatively new. One of our findings is that these methods functioned well with data as high as five dimensions. Investigating item parameter recovery, the results do suggest some practical and important findings. It was stated that using broad sample sizes and test lengths provided more accurate results than using the other test conditions that included smaller sample sizes and test length. This finding is consistent with the research carried out by Reise and Yu (1990); Lautenschlager, Meade, and Kim (2006); Bahry (2012); Montgomery and Skorupski (2012); Wang and Chen (2005); DeMars (2002); Hulin, Lissak, and Drasgow (1982); and J. Lee (2012).

In the present study, MCMC estimation was performed using 8,000 iterations, and not surprisingly, for all the test conditions, the highest estimation time was required for the MCMC technique. The estimation time of parameters by using BA-EM was shorter than MCMC and MH-RM. In addition to these results, the estimation time of parameters that was estimated by MH-RM was higher than MCMC. When the estimation time was compared, one of the important results is that using BA-EM is very practical. Another important situation is software availability. flexMIRT is not free but researchers can obtain BMIRT easily, because it is free. Furthermore, all these results were confirmed with multidimensional dichotomous and polytomous models by using different test structures and different test conditions.

Determining the priors that were used for estimation is among the important limitations of this study. Another limitation is the computing power that is used to estimate the parameters. Especially estimation of data sets that include larger sample size and test length required a computer that had better computation capacity. Another

limitation is deciding on the iteration number. If researchers have not enough time, BA-EM may be more practical than MCMC and MH-RM. In addition, if researchers are interested in using MCMC, a higher iteration number can provide more satisfying results.

In summary, each of the software packages had both strengths and weaknesses, and users need to make their choices carefully. The test conditions such as sample size, test length, number of dimensions, and the correlation between dimensions have an important effect on estimation. Larger sample sizes and test lengths can provide more accurate results. If data sets include higher dimensions such as five, the computing power is important for estimation. MCMC and MH-RM can provide better results. If the data sets consist of fewer dimensions and if the model is not complicated, BA-EM, MCMC, and MH-RM should be preferred. Especially, BA-EM is very practical because of the estimation time. If researchers do not have the funding but they have enough time, MCMC (BMIRT) is the best option. If data sets have complex conditions such as larger sample size and test length, BMIRT (MCMC) or flexMIRT (MH-RM) should be preferred. For analyses, writing command by using BMIRT is harder than flexMIRT. In other words, flexMIRT is easy to use. Therefore, if users are not familiar with the DOS commands that is used in BMIRT, then flexMIRT should be preferred.

Authors' Note

In this study, the bias values and RMSE values were computed for all test conditions, but because our article will exceed the page limit, all other results can be provided via e-mail. The average estimation times of a few test conditions are given in the article. If researchers are interested in all other results, they may ask for them via e-mail (gyavuz2010@gmail.com). This paper is based on the part of my Ph.D dissertation. This article was presented at the meeting of NCME, Chicago, 2015.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size* (Master's thesis). University of Alberta, Edmonton, Canada. Retrieved from <http://hdl.handle.net/10402/era.25356>

- Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541-562.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 29, 395-414.
- Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis. Retrieved from http://conservancy.umn.edu/bitstream/handle/11299/155592/Bulut_umn_0130E_13879.pdf?sequence=1
- Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75, 581-612.
- Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.
- Cai, L. (2010c). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.
- Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Lincoln, IL: Scientific Software International.
- Cai, L., & Monroe, S. (2013). IRT model fit evaluation from theory to practice: Progress and some unanswered questions. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 102-106.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, 1-29.
- De la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, 33, 465-485.
- De la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, 35, 296-316.
- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education*, 15, 15-31.
- Finch, W. H. (2011). Multidimensional item response theory parameter estimation with non-simple structure items. *Applied Psychological Measurement*, 35, 67-82.
- Finch, W. H., & French, B. F. (2012). A comparison of methods for estimating confidence intervals for omega squared effect size. *Educational and Psychological Measurement*, 72, 68-77.
- Guyer, R., & Thompson, N. A. (2011). *User's manual for Xcalibre item response theory calibration software* (Version 4.1.3). St. Paul, MN: Assessment Systems Corporation.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31, 457-459.
- Hsieh, M., Proctor, T. P., Hou, J., & Teo, K. S. (2010). A comparison of Bayesian MCMC and marginal maximum likelihood methods in estimating the item parameters of the 2PL IRT model. *International Journal of Innovative Management, Information & Production*, 1(1), 81-89.
- Hulin, C. L., Lissak, R. L., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6, 249-260.

- Jurich, D., & Goodman, J. (2009, October). *A comparison of IRT parameter recovery in mixed format examinations using PARSCALE and ICL*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.
- Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*, 399-419.
- Lautenschlager, G. J., Meade, A. W., & Kim, S.-H. (2006). *Cautions regarding sample characteristics when using the graded response model*. Paper presented at the conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Lee, J. (2012). *Multidimensional item response theory: An investigation of interaction effects between factors on item parameter recovery using Markov chain Monte Carlo estimation* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI. Retrieved from <http://etd.lib.msu.edu/islandora/object/etd%3A1577>
- Lee, S. H. (2007). *Multidimensional item response theory: A SAS MDIRT MACRO and empirical study of PLAT MATH Test* (Unpublished doctoral dissertation). The University of Oklahoma, Norman.
- Montgomery, M., & Skorupski, W., (2012). *Investigation of IRT parameter recovery and classification accuracy in mixed format*. Paper presented at the meeting of the National Council of Measurement in Education, Vancouver, Canada.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.
- Svetina, D. (2011). *Assessing dimensionality in complex data structures: A performance comparison of DETECT and NOHARM methods* (Unpublished doctoral dissertation). Arizona State University, Tempe, AZ.
- Wang, W. C., & Chen, C. T. (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.
- Weiss, J. D., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG* (Technical Report). St. Paul, MN: Assessment Systems Corporation.
- Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory*. Monterey, CA: CTB/McGraw-Hill.
- Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83-105.
- Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30*, 469-492.
- Zhang, B., & Stone, C. A. (2004, April). *Direct and indirect estimation of three-parameter multidimensional item response models*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.
- Zhang, B., & Stone, C. A. (2007). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*, 181-196.
- Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*, 375-398.