Spring 5-12-2017

Comparison of IRTPRO 3 and Mplus 7 for multidimensional item response item parameter and examinee ability estimation

Tianna Sims

ACCEPTANCE

This dissertation, A COMPARISON OF IRTPRO 3 AND MPLUS 7 FOR MULTIDIMENSIONAL ITEM RESPONSE ITEM PARAMETER AND EXAMINEE ABILITY ESTIMATION, by TIANNA CHANTEL SIMS, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education and Human Development, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

T. Chris Oshima, Ph.D.
Committee Chair

Hongli Li, Ph.D.
Committee Member

Tim Sass, Ph.D.
Committee Member

Nicole Patton Terry, Ph.D.
Committee Member

Keith Wright, Ph.D.
Committee Member

Date

William Curlette, Ph.D.
Chairperson, Department of Educational Policy Studies

Paul A. Alberto, Ph.D.
Dean, College of Education and
Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education and Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

____

TIANNA CHANTEL SIMS

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Tianna Chantel Sims
Educational Policy Studies
College of Education and Human Development
Georgia State University

The director of this dissertation is:

T. Chris Oshima
Department of Educational Policy Studies
College of Education and Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

Tianna Chantel Sims

ADDRESS: Georgia State University
College of Education and Human Development
Department of Educational Policy Studies
30 Pryor Street Suite 450
Atlanta, GA 30302

EDUCATION:

Ph.D.	2017	Georgia State University Educational Policy Studies
Masters Degree	2007	Florida International University Public Heath
Bachelors Degree	2004	University of Central Florida Math/Engineering

PROFESSIONAL EXPERIENCE:

2015-present	Accountability specialist Georgia Department of Education
2012-2015	Project Manager Urban Child Study Center
2008-2012	Teacher Gwinnett County Schools

SELECTED PRESENTATIONS AND PUBLICATIONS:

Floyd, T.C.S., Towson, J., Bingham, G., & Patton Terry, N., (2015) *Classroom Assessment Scoring System in Diverse Settings: Confirmatory Factor Analysis*. Paper presented at the National Council on Measurement in Education Conference

Floyd, T.C.S., Towson, J., Bingham, G., & Patton Terry, N., (2015) *PreK Teachers' Language Practices: Examining the Effect on Student Language & Pre-Literacy Skills*. Paper presented at the American Educational Research Association Conference

Towson, J., **Floyd, T.C.S.,** Bingham, G., & Patton Terry, N., (2014) *Providing Children High-Quality Language Modeling: Instructional Support Construct of the Classroom Assessment Scoring System*. Paper presented at the American Speech Language and Hearing Association

Marencin, N., **Floyd, T.C.S,** & Patton-Terry, N. (2013) *Measurement Differences in Performance of Non-mainstream American English Speaking Pre-schoolers on the CELF-P* . Poster presented at the American Speech Language and Hearing Association Conference

Floyd, T.C.S., Bakhtiari, D., Patton-Terry, N., & Miti Mansour, S., (2013). *Contributions of Oral Language, Home Literacy Environment, and Parental Education to Diverse Learner's Emergent Literacy*. Paper presented at the Georgia Educational Research Association Conference

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2012-2016	American Educational Research Association
2013-2016	National Council on Measurement in Education
2016	Psychometric Society

A COMPARISON OF IRTPRO 3 AND MPLUS 7 FOR MULTIDIMENSIONAL ITEM RESPONSE ITEM PARAMETER AND EXAMINEE ABILITY ESTIMATION

by

TIANNA CHANTEL SIMS

Under the Direction of T. Chris Oshima

ABSTRACT

Advances in computational ability have produced new software, such as Mplus 7 (Muthén & Muthén, 2011) and IRTPRO 3 (Cai, Thissen, & du Toit, 2015), which can estimate multidimensional item parameters and examinee abilities. Due to its nascence, little research has been done on the ability of IRTPRO 3 to estimate multidimensional item parameters and examinee abilities in comparison to other available software. This study investigates the capability of Mplus 7 and IRTPRO 3 to recover multidimensional item parameter and examinee ability levels under different conditions, including estimation techniques, test lengths, sample sizes, correlations and test structures. The results show that certain estimation techniques available in each software package work well under the given conditions. There are, however, some instances where each estimation techniques encountered challenges. Practitioners should take into consideration the test model and the information needed when selecting the most appropriate software and estimation technique.

INDEX WORDS: Multidimensional IRT, simulation, estimation

A COMPARISON OF IRTPRO 3 AND MPLUS 7 FOR MULTIDIMENSIONAL ITEM RESPONSE ITEM PARAMETER AND EXAMINEE ABILITY ESTIMATION

by

TIANN CHANTEL SIMS

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Educational Policy Studies

in

Educational Policy Studies

in the

College of Education and Human Development

Georgia State University

Atlanta, GA
2017

DEDICATION

This dissertaion is dedicated to my parents. Although my father did not say it often, I knew he was proud of me and loved me strongly. My mother has always supported me, loved me, and has shown me how to continuously look for the postives life has to offer. To my siblings Antoine, Jonathan, Brandon, Devonna, and Donovan, thanks for all of the support throughout the years. Thanks to the rest of my family, aunts, uncles, grandparents, and cousins for believing in me from the beginning. A special note goes out to my best friend Raelyn who would indulge me in middle school with " paging Dr. Sims".

I would also like to thank all of the wonderful people that I have met through the process who have encouraged me to keep on going. Thanks to my wonderful and support of friends in Atlanta whose shoulders I have cried on and who have seen me grow through this process. Thank you to my doctors whose medical guidance helped me to see this day and to Beth whose counsel kept me going through the dark times in my life. Thank you Dr. Patton Terry for taking a chance on me from the very beginning. Finally, a true note of gratitude goes out to my advisor Dr. Chris Oshima. Thank you for not only being an advisor but a true friend who stood besides me through all of the ups and downs.

ACKNOWLEDGMENTS

I woud like to acknowledge the Southern Regional Education Board for selecting me to become one of their docortal fellows and for the accompanying funding that allowed me to complete my program. I would also like to acknowledge the College of Education and Human Development for awarding me a Doctoral Dissertation Grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1PL	One parameter Logistic
2PL	Two parameter logistic
3PL	Three parameter logistic
ADQ	Adaptive quadrature
CTT	Classical test theory
EAP	Expected a posteriori
FA	Factor analysis
ICS	Item characteristic surface
IRT	Item response theory
MAP	Maximum a posteriori
MDIFF	Multidimensional difficulty
MDISC	Multidimensional discrimination
MHRM	Metropolis-Hastings Robinson-Monro
MIRT	Multidimensional item response theory
ML	Maximum likelihood
UIRT	Unidimensional item response theory
WLSMV	weighted least squares mean and variance

1 THE PROBLEM

Testing has been and will always be an integral part of the world of education. There are numerous approaches that attempt to take information from a test and convert it to information that can be used to understand student ability. The advancements in technology have led to the development of computational techniques and software that can turn test information into ability information. The increase in computation techniques first benefited the field of classical test theory (CTT) by making calculations, such as point biserial correlation coefficient and item discrimination indices, easier. Computational advances continued to benefit the area of item response theory (IRT) specifically, unidimensional item response theory (UIRT). Programs such as BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2002) and WINSTEP (Linacre, 2010) were developed to help researchers more accurately model examinee ability. Finally, computational advances improved enough to benefit multidimensional item response theory (MIRT). New software, such as Mplus 7 (Muthén & Muthén, 2011), flexMIRT (Cai & Wirth, 2013) and IRTPRO 3 (Cai, Thissen, & du Toit, 2015a) have been developed and allow for estimation of MIRT models.

The increase in technology has let to numerous choices when it comes to estimating multidimensional models. It is important for educational professionals to understand the differences among the available software. Equally important is the understanding of how accurately the software can estimate models, the default settings, and model estimation techniques the software employs. Both IRTPRO 3 and Mplus 7 approach multidimensional models in a different fashion. IRTPRO 3 approaches models from the item response theory framework while Mplus 7 approaches models from the factor analysis framework.

Research Questions

Given the new techniques available in multidimensional test modeling and the necessity to make sure software can estimate item parameters and examine abilities even in the most complex structure, this study will answer the following questions: (1) How do Mplus 7 and IRTPRO 3 differ in their estimation techniques? (2) How do Mplus 7 and IRTPRO 3 compare in the recovery of multidimensional item parameter estimates under differing conditions?, and (3) How do Mplus 7 and IRTPRO 3 compare in the recovery of examinee ability estimates under differing conditions?

Significance

There have been multiple studies that investigated the use of MIRT software such as Mplus 7, flexMIRT, and TESTFACT (Wilson, Wood & Gibbons, 1991). Many of these studies incorporated a variety of simulation conditions, including changes in the number of examinees, changes in the size of the correlation between dimensions, and changes in the number of items tested. Many of these studies though, have focused on the two-dimensional test structure. In addition, there has been a lack of research on the estimation techniques contained in IRTPRO 3 because of the nascence of the program. An article by Han and Paek (2014) looked at the ability of IRTPRO 3 and Mplus 7 to recover item parameters, however, the authors did not compare the capability of the programs to estimate examinee ability nor did the study vary simulation conditions. The study only evaluated four different multidimensional test structures and simulated data for 3,000 examinees on a 30-item test.

The present study will include an investigation of the ability of each type of software to recover examinee ability as well as expand the number of simulation conditions to inclusively assess each software’s estimation abilities. Additionally, previous studies have focused on theory and not on practitioners. This study takes a unique perspective in that it focuses on practitioners who might want to include psychometric properties in their test construction yet do not have the resources for a full psychometric analysis. These practitioners include school districts and large schools that might want to start a small psychometric or test development department in order to create common district or school level assessment. Additionally, those working in credentialing, licensing and other types of certification granting organizations might also benefit from the results of this study.

2 REVIEW OF THE LITERATURE

IRT is a method used to ascertain an examinee’s ability level and uses item parameters to estimate ability levels. IRT is different from CTT in a variety of ways. The first difference is that CTT is both test dependent and sample dependent. Test dependency means that an examinee’s ability is related to the difficulty of the test. For example, on an easy test, an examinee’s ability might appear high and on a difficult test, their ability might appear low. CTT is also sample dependent and IRT is not. Sample dependency means that the test scores are meaningful when the examinee who takes the test is similar to the population for which the test was made. In item response theory, item parameters and examinee abilities estimates do not depend on each other, as they do in CTT.

Unidimensional Item Response Theory

IRT postulates that an examinee’s ability can be explained by underlying traits. UIRT assumes that there is only one underlying trait that explains the probability of the correct answer given the item parameters. For example, a math test is only testing math ability; it is not also testing reading or science ability. Another assumption of item response models is that ability and probability of a correct response is a monotonically increasing relationship. In addition, UIRT assumes local independence. The assumption of local independence means that no item can affect another item. For instance, an examinees response on question 12 does not affect their response on question 17.

IRT models are written as probabilistic functions where the probability of correctly answering a question is a function of the examinee’s ability level, item discrimination, item difficulty and a guessing parameter. The general dichotomous UIRT model is shown below where θ is the examinee’s ability, *a* is the item discrimination, *b* is the items difficulty and *c* is the pseudo guessing parameter.

$$P\big(x_{ij} = 1|\theta_j, a_i, b_i, c_i\big) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}} \qquad (1)$$

The equation shown above is often referred to as the three-parameter logistic model (3PL). The two-parameter logic model (2PL) is a special case of the 3PL model where c = 0 and the one-parameter logistic model (1PL) is the simplest of UIRT models where c = 0 and a = 1.

There are a few extensions of these three basic UIRT models including polytomous response models. The polytomous item response model handles items that can take on more than two values. Polytomous models include one, two, or three parameters models as well as the rating scale model, the partial credit model and the graded response model.

Multidimensional Item Response Theory

Multidimensional item response theory (MIRT) is an extension of UIRT. Where UIRT assumes unidimensionality, MIRT does not. MIRT addresses the assumption of local independence in the basic IRT models. These models account to the fact that items can test more than one ability or trait. Multidimensional models can separate the underlying abilities and, theoretically, there is no limit to the number of abilities that can be assessed.

Researchers started working in the late 1970s and early 1980s to develop MIRT models. Before MIRT models were introduced, scientists first had to conceptualize them. Lord and Novick (1968) worked on conceptualizing and developing MIRT models. Although never coming up with a full model, Lord and Novick conceptualized MIRT including latent space and local independence (Reckase, 2009). Samejima also conceptualized and tried to model MIRT. Her model used continuous responses and not dichotomous responses (Reckase, 2009).

Unfortunately, Samejima’s model was not widely employed and only found use in one application.

After models were conceptualized, researchers started developing mathematical MIRT models. Muliak (1972) came up with a model that changed as the number of dimensions increased. When the probability of a correct response was fixed, the values of the exponents in the equation would change as the number of dimensions increased (Reckase, 2009). Muliak’s model led the way to the models of Sympson (Sympson, 1978) and Whitely (Whitely, 1980). Both of these research scientists proposed a model in which the scaling of the item parameters changed with the change in the number of dimensions. Nonetheless, none of the aforementioned models were full MIRT models because the probability of a response changed when the number of dimensions changed. Finally, Reckase and McKinley (1991) proposed a model in which the probability of a correct response stayed the same and did not increase or decrease with the number of dimensions.

It was through the research of Reckase and McKinley (1991) that the MIRT model that is often used today was developed. Currently there are two types of MIRT models, the compensatory and the noncompensatory model. There are also MIRT polytomous models, graded response models, and partial credit models (De Ajaya 2013; Reckase 2009).

Compensatory model. The basic MIRT equation relates the probability of an examinee correctly answering an item to the examinee’s ability level. The compensatory models are as follows; where $\boldsymbol{\theta}$ is a vector of abilities, $\boldsymbol{a}$ is a vector of item discrimination parameters and $\boldsymbol{b}$ is the difficulty term.

$$P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{e^{a_i(\theta'_j - b_i)}}{1+e^{a_i(\theta'_j - b_i)}} \quad (2)$$

The compensatory model allows an examinees ability in one area to compensate for the ability in another area. This model features the additive property of logit. For example, if a student has high algebra ability, that might compensate for lower geometry ability. Another way to express the compensatory model is in slope intercept form were d_i is t equal to $-\boldsymbol{a_i b_i}$.

$$P(x_{ij} = 1|\theta_j, a_i, d_i) = \frac{e^{a_i\theta'_j + d_i}}{1+e^{a_i\theta'_j + d_i}} \quad (3)$$

In the compensatory model, each ability dimension has a separate *a* parameter for the item, yet, there is only one *d* parameter. Using a two-dimensional model as an example, the graphical representation an examinee correctly answering a question is a function of the difference between the items *i*'s location and the examinee *j*'s ability on the first dimension and items *i*'s location and the examinee *j*'s ability on the second dimension. Instead of producing an item characteristic curve as in the case of UIRT, MIRT models produce an item characteristic surface (ICS). For the two-dimensional model, the axes of the ICS are θ_1, θ_2, and the probability of a correct response. Figures 1 and 2 show the ICS for two compensatory items.

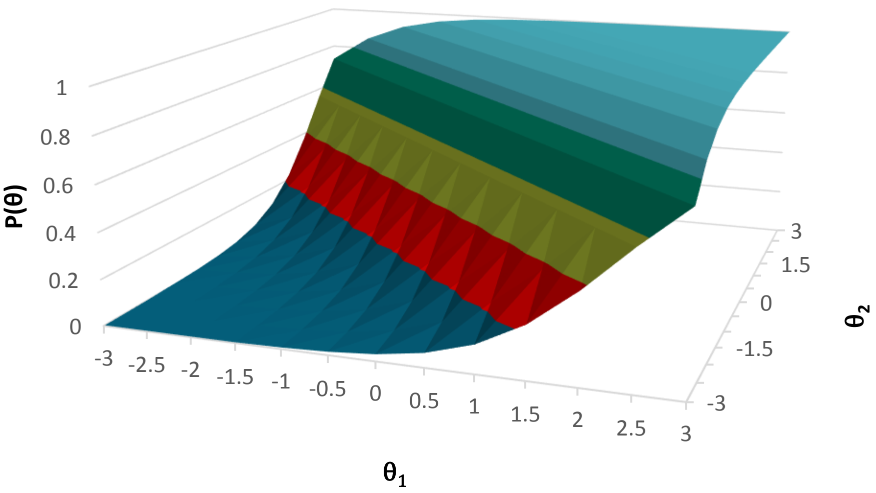

Figure 1. Compensatory ICS a_1 = 1.5, a_2 = 1.5 d = 1.5

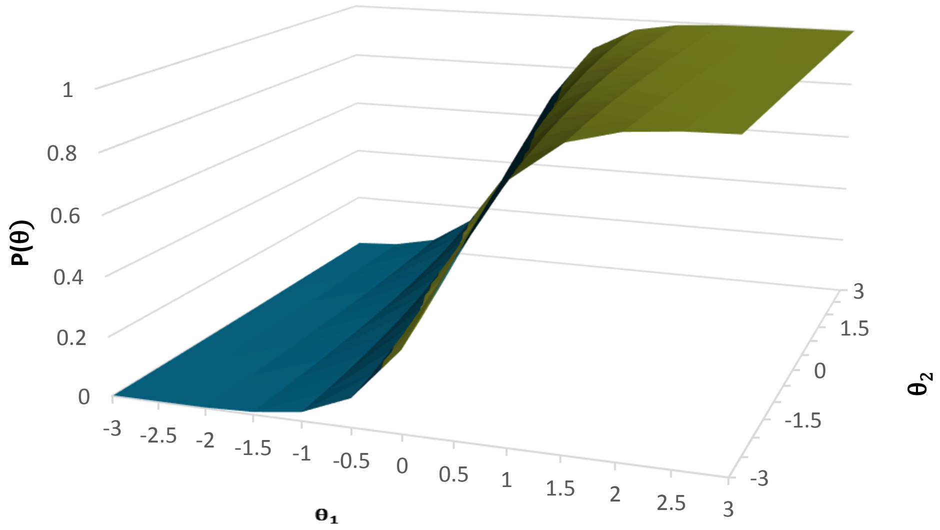

Figure 2. Compensatory ICS a_1 = 0.5, a_2 = 2.5 d = 0.5

Compensatory multidimensional models can also incorporate discrimination and difficulty. These concepts are similar to what is encountered in UIRT; however, the added dimensions make it a more intricate model. As described previously, ***a*** is a vectors of slopes for the item where there is one slope per dimension assessed. Aggregating information about the ***a*** parameter for all dimensions assessed produces the multidimensional discrimination parameter (MDISC) where *m* is the number of dimensions. MDISC can also be used to calculate multidimensional item difficulty (MDIFF). MDIFF takes into account the multiple ***a*** parameters that are available for a given item and uses them to calculate the item's difficulty. Both MDISC and MDIFF are interpreted in a similar fashion as the *a* and *b* UIRT parameters.

$$\text{MDISC} = \sqrt{\sum_{v=1}^{m} a_v^2} \qquad (4)$$

$$\text{MDIFF} = \frac{-d}{\sqrt{\sum_{v=1}^{m} a_v^2}} \quad (5)$$

Noncompensatory model. The noncompensatory model does not allow the ability in one area to compensate for the ability in another area. The noncompensatory model features the multiplicative property of logit. For example, a student's high math ability, might not effect reading comprehension at all. In the noncompensatory model, each dimension will have a separate *a* value and *b* value and a separate probability of a correct response. The overall probability of correctly answering the question is the product of the two separate probabilities. The probability of correctly responding to an item is given in the equation below. Noncompensatory models also produce ICS. Figures 3 and 4 show two ICS from noncompensatory MIRT items.

$$P\left(x_{ij} = 1 | \theta_j, a_i, b_i, c_i\right) = c_i(1 + c_i) \prod_{m=1}^{m} \frac{e^{a_{im}(\theta_{jm} - b_{im})}}{1 + e^{a_{im}(\theta_{jm} - b_{im})}} \quad (6)$$

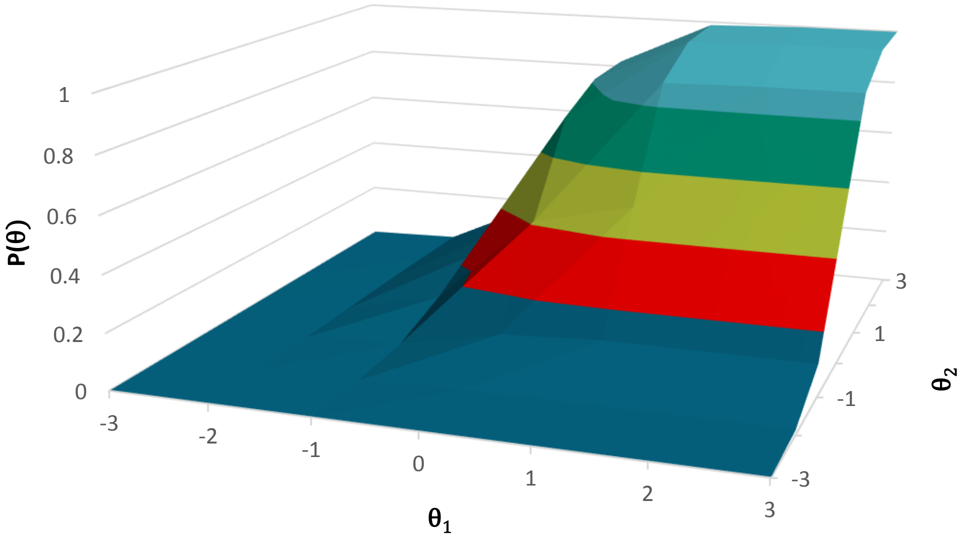

Figure 3. Noncompensatory ICS a_1 = 1.8, a_2 = 1.2 b_1 = -0.5, b_2 = 0, *c* = 0

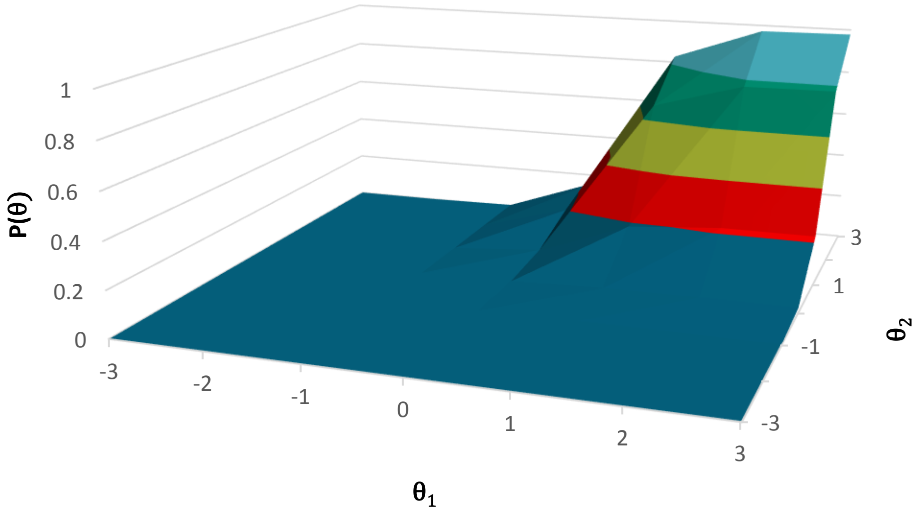

Figure 4. Noncompensatory ICS a_1 = 2.0, a_2 = 1.5 b_1 = 0.5, b_2 = 1.5, c = 0

In MIRT, there are different underlying structures that show how the abilities related to the different items on the test. In a between-item multidimensional model, there are multiple abilities being tested, but, each item only test one ability. In the within-item model, one item can test multiple abilities. The entire test structure can also be a combination of the between-item model and the within-item model.

Factor Analysis

Factor analysis (FA) is similar to IRT and is often used in the case of continuous underlying variables. In factor analysis, observed variables are used to identify underlying traits or variables. These latent variables explain how the observed variables are correlated with each other. Factor analysis based on matrix algebra uses either a correlation matrix or a covariance matrix. The correlation matrix produces standardized data and the covariance matrix does not.

When variables are dichotomous as in the case of test data, factor analysis responses are thought to be controlled by and underlying or latent variable, Y and a threshold τ. The underlying latent variable is similar to the underlying ability on an IRT model and the threshold, τ, is similar to the IRT difficulty parameter. An examinee's response then can be modeled by equation 7 and 8, where **Y** is the vector of latent underlying variables, **Λ** is a matrix of factor loadings, **θ** is the latent variable vector of ability and **E** is a vector of residuals. The assumptions of FA are that **θ** has a multivariate normal distribution with a mean of zero and a standard deviation of one, and **E** is multivariate normal with a mean of zero and a standard deviation of ψ^2, where ψ^2 is the diagonal matrix with positive elements (Knol & Berger 1991, Muthén 1978).

$$\boldsymbol{X}_i \begin{cases} 1, if\ \boldsymbol{Y}_i \geq \ \boldsymbol{\tau}_i \\ 0, otherwise \end{cases} \tag{7}$$

$$\mathbf{Y}=\mathbf{\Lambda}\theta+\mathbf{E} \tag{8}$$

Item Response Theory and Factor Analysis

IRT and FA differ mathematically in how examinee abilities and item parameters are calculated as well as in their model fit statistics. For IRT, model fit is relatively straight forward for the 1PL model, however, it can get a bit more complicated with the 2PL and 3PL models (Orlando & Thissen 2003). In order to get model fit statistics for the 2PL and 3PL models, examinees are first sorted by ability and subgroups are formed based on the sorting. Next, the proportion of examinee's correct and incorrect responses are calculated based on subgroups, and lastly the observed proportions are compared to the predicted responses (Bock, 1972; McKinley & Mills, 1985; Orlando & Thissen, 2000; Orlando & Thissen 2003; Yen, 1981). Yen (1981) developed a Q_1 measure which first separates ability into 10 levels and then constructs a 10 x 2 contingency table based on responses. Finally, the expected proportion is calculated from the model based on the mean predicted probability. Bock (1972) developed a χ^2 measure similar to the Q_1. Bock's χ^2 measure varies the number of interval and the measure uses the median predicted probability as opposed to the mean. Others have also proposed χ^2 measures similar to the ones proposed by Yen and Bock including Wright and Mead (1977) χ^2 and McKinley & Mills (1985) G^2.

Orlando and Thissen (2000 and 2003) expanded on the previous IRT model fit work and proposed two new indices $S - X_i{}^2$ and $S - G_1^2$. They advise that the previous calculations are problematic because the degrees of freedom are hard to identify and placing examinees into intervals is decidedly sample dependent which therefore influences the subsequent cut points and the overall fit statistic. Their proposed measures are based on the observed and expected frequencies correct and incorrect as opposed to only ability. The expected frequencies are calculated using a model predicted joint likelihood distribution. The results show that the $S - X_i{}^2$ measure could be useful in identifying model fit in a variety of IRT models including 2PL and 3PL models (Orlando and Thissen 2003).

Factor analysis also has a variety of fit indices. The χ^2 goodness of fit statistic compares the sample to the covariance matrix that was used to fit the model (Hu & Bentler 1999). This index is a function of sample size and therefore produces results that indicate good model fit when the sample size is large. Indices that compare the fit of nested models are based on the goodness of fit statistics. One index is called the normed fit index (NFI) (James, Mulaik & Brett 1982). This index compares the goodness of fit statistics for nested models. Bentler and Bonnett (1980) developed the nonnormed fit index (NNFI) which takes into account the sample size. The comparative fit index (CFI) is related to the NFI and also takes into consideration sample size thus making it good even for smaller sample sizes (Bentler 1990). Other measures of model fit are the Tucker Lewis index (TLI;1973) and Bollen's (1989) incremental fir index (IFI). Taken together, these indices are often referred to as incremental fit measure because they measure the improvement of one model over a more restricted or baseline model (Hu & Bentler 1999).

Another class of fit indices are absolute fit measures. These measures include the goodness of fit index (GFI), the adjusted goodness of fit index (AGFI), root mean residual (RMR), standardized root mean residual (SRMR) and the root mean square error of approximation (RMSEA). Jöreskog and Sörbom’s GFI (Jöreskog & Sörbom, 1984) calculates the proportion of variance explained and uses the estimated population covariance. The AGFI (Bentler, 1983; Tanaka & Huba, 1985) adjust the GFI by adding in a component of degrees of freedom. The AGI also favors more parsimonious models over more complex models. Both the RMR (Jöreskog and Sörbom 1981) and the SRMR (Bentler 1995) measure the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance matrix. The difference between the two indices is that the RMR reports results based on the scale of the variables whereas the SRMR standardizes the information. The RMSEA developed by Steiger (1990) compares the estimated model to the population covariance matrix. Confidence intervals can be calculated for the RMSEA giving it an advantage over other fit indices (Hooper, Couglan and Mullen, 2008; MacCallum et all 1996).

Estimates from IRT and FA can be formulated such that they are comparable. The equations for transforming IRT estimates to FA estimates are shown in equations 9 and 10 where λ_i is the i^{th} row of the $\mathbf{\Lambda}$ vector, φ is the covariance matrix of factors, and the other variables have been defined previously (De Ayala 2013, Finch 2010, Kamata & Bauer 2008, and McDonald 1999).

$$a_i = \frac{\lambda_i}{\sqrt{(1-\grave{\lambda}_i\varphi\lambda_i)}} \qquad (9)$$

$$b_i = \frac{-\tau_i}{\sqrt{(1-\grave{\lambda}_i \varphi \lambda_i)}} \tag{10}$$

Parameter Estimation

Often times neither the examinee's ability nor the item parameter values are known. In cases, such as these, it is necessary to estimate both the ability and item parameters at the same time. The estimation of both item parameters and examinee ability simultaneously is referred to as joint estimation. There are multiple methods used to estimate both item and examinee parameters. Often, the estimation procedure used is related to the software used for the analysis. As mentioned in the introduction, there are numerous software platforms available to estimate MIRT models. The estimation procedures elucidated in this literature review corresponded to the estimation procedures that are in IRTPRO 3 and Mplus 7 as these are the two programs that are the focus of this study. IRTPRO 3 offers the Bock-Aitkin (BA) approach with expectation maximizing algorithm (Bock & Aitkin, 1981) Adaptive quadrature (ADQ) (Schilling & Bock, 2005) and Metropolis-Hastings Robbins-Monro (MHRM) (Cai, 2010a, Cai, 2010b). Mplus 7 offers a variety of estimation techniques, however, this study will use weighted least squares (WLSMV), Bayes, and maximum likelihood (ML) (Muthén, 1978; Muthén, 1984).

Bock-Aitkin. The BA approach is also known as marginal maximum likelihood approach. In this approach, the probability of obtaining a specific item response pattern in a population of examinees is calculated by "weighting the likelihood by the probability density of the **θ** vector and then integrating over the **θ** space" (Reckase, 2009). The formula for this approach is shown below where $P(u = u_l)$ is the probability of the particular score pattern for the entire population of examinees, $L(u_l|\theta)$ is the likelihood of the particular item response pattern for a particular ability vector, and $g(\theta)$ is the θ probability density function.

$$P(u = u_l) = \int_\theta L(u_l|\theta)g(\theta)d\theta \quad (11)$$

Adaptive quadrature. ADQ is an extension of the BA approach and can be used when a high number of dimensions are present. This approach estimates the probability of a certain response pattern in the population by using the multivariate density function. The ADQ approach does not use examinee's ability, it utilizes quadrature points and only item parameters can be estimated. The equations for the ADQ probability function and the quadrature points are shown below where *N* is the number of examinees, *s* is the number of items, r_w is the frequency of item response pattern *w*, X is the vector of values for the quadrature point, q is the number of quadratures, *k* is the dimension, and $A(X_{qk})$ relates to the height of the normal density function and the corresponding weight related to the height of the function (Reckase, 2009, Schilling & Bock, 2005). All other variables are defined previously.

Probability function

$$L(U) = \frac{N!}{r_1!r_2!\ldots r_s!}P(u = u_1)^{r_1}P(u = u_2)^{r_2}\ldots P(u = u_s)^{r_s} \quad (12)$$

Quadrature integration

$$P(u = u_l) = \sum_{qm=1}^{Q}\cdots\sum_{q2=1}^{Q}\sum_{q1=1}^{Q}L(u_l|X)A\left(X_{q1}\right)A\left(X_{q2}\right)\ldots A\left(X_{qm}\right) \quad (13)$$

Metropolis-Hastings Robbins-Monro. The MHRM estimation algorithm is implemented in a three-step process, which uses random imputations from the Metropolis-Hastings sampler (Wang & Nydick, 2015). First, there is stochastic imputation, next stochastic approximation and finally a Robbins-Monro update. The three steps are implemented for every iteration of the algorithm, and the algorithm stops when the model converges.

Weighted least squares. The WLSMV approach in Mplus 7 attempts to minimize the sum of squares of the difference between the vector of all latent response variables correlations and threshold estimated from the sample (s) and the relating population correlations (σ) (Muthén & Asparouhov, 2013). The weighted least squares approach uses the following equation where diag(W)$^{-1}$ is the diagonal of the weight matrix (Muthén & Asparouhov, 2013).

$$F = (s - \sigma)' diag(W)^{-1}(s - \sigma) \tag{14}$$

Bayes. Bayesian estimation is based on a theorem that explains the relationship between an event and the possible causes. In IRT, the event is item response pattern and the possible causes are the possible ability levels (Reckase, 2009). Just as in ML estimation, the ability vectors that are chosen in Bayesian estimation are the highest possible ability vector that would produce the particular item response pattern. When using Bayesian estimation in Mplus 7, the default is to use non-informative priors. The equation for Bayesian estimation is shown below where $h(\theta|U_j)$ is the posterior probability density of θ for the given the item response pattern and $f(\theta)$ is the prior probability density function for θ. All other variables are defined in the preceding equations.

$$h(\theta|U_j) = \frac{L(U_j|\theta)f(\theta)}{\int_\theta L(U_j|\theta)f(\theta)d\theta} \tag{15}$$

Ability Estimation

Both IRTPRO 3 and Mplus 7 can estimate item parameters and examinee ability. IRTPRO 3 uses Maximum a posteriori (MAP) and expected a posteriori (EAP) while Mplus 7 uses ML and Bayes.

Maximum likelihood. The ML estimation procedure takes an examinee's response pattern and finds the highest ability vectors that correspond to the highest probability of selecting that particular response patter. Maximum likelihood estimation is the product of the probability of an examinee answering the first item as seen in the response pattern, times the probability of the examinee answering the second item as is seen in the response pattern and so on. The equation for ML estimation in MIRT is shown below where $L(U_j|\theta_j)$ is the likelihood of a person with ability *j* located at $\boldsymbol{\theta}_j$ answering in the response pattern ***U****j*, *Q* is *(1-P)* and *u* is the item score. All other variables are defined in the preceding equations.

$$L\left(U_j\middle|\theta_j\right) = \prod_{i=1}^{n} P(u_{ij}|\theta_j)^{u_{ij}} Q(u_{ij}|\theta_j)^{1-u_{ij}} \tag{16}$$

Maximum a posteriori. The maximum a posteriori (MAP) estimation technique is similar to maximum likelihood in that the entire scoring pattern is used to produce the ability estimate. Be that as it may, MAP is Bayesian estimator, and ML is not. In addition to the maximum likelihood function, the MAP uses the assumed population distribution to estimate ability. The MAP is also a Bayesian estimation technique, but, it uses the assumed population distribution along with the likelihood function.

Expected a posteriori. The EAP estimation technique is an extension of the MAP. Rather than finding the maximum point using MAP, the average value with the weighted function is used to calculated the ability estimate. The EAP estimation also address the issue of student examinees who answer all questions correctly or incorrectly. The equation for EAP is shown in Equation 17, the posterior standard deviation is shown in Equation 18, and variables have been previously defined.

Equation 17 EAP estimation

$$\hat{\theta}_j = \frac{\sum_{q=1}^{q} X_q L(X_q) A(X_q)}{\sum_{q=1}^{q} L(X_q) A(X_q)} \qquad (17)$$

Equation 18 EAP posterior standard deviation

$$PSD(\hat{\theta}) = \sqrt{\frac{\sum_{q=1}^{q} (X_q - \widehat{\theta_j})^2 L(X_q) A(X_{q)}}{\sum_{q=1}^{q} L(X_q) A(X_q)}} \qquad (18)$$

where

$$L(X_q) = \prod^{n} p_j(X_q)^{X_{ij}} (1 - p_j(X_q))^{(1-x_{ij})} \qquad (19)$$

One noted differences between the EAP and MAP estimation technique is that the MAP uses an iterative approach and the EAP is based on the quadrature method. Additionally, MAP uses a continuous distribution and EAP uses a discrete distribution. Finally, EAP is based on the mean and MAP is based on the mode.

Previous comparison studies. As stated previous, there have been other studies that have investigated item parameter recovery and examinee ability recovery under various test models and other simulation conditions. One study was conducted by Knol and Berger (1991). In their study, they compared IRT to factor analysis using TESTFACT (Wilson, Wood & Gibbons, 1984), NOHARM II (Fraser & McDonald 1988), and MAXLOG (McKinley & Reckase, 1983) for the IRT approach. For the factor analysis approach, they used SPSS and the iterative principal factor analysis (IPFA), unweighted least squares (ULS), and GLS common factor analysis (GLS) and alpha factor analysis (ALPHA). They also used LISREL VI (Jöreskog & Söbom, 1984) and the minimum residuals factor analysis method (MINRES). Due to the computational limitations, they only employed four test models. Each model included one to three dimensions, either 15 or 30 items, and three sample sizes, 250, 500 and 1000 examinees. Knol and Berger compared results using the mean squared difference and found that MAXLOG performed the worst for the two-dimensional model with a mean squared error almost double what was seen with the other estimation techniques. For the three-dimensional model with 30 items, NOHARM and the factor analysis approaches performed better than both TESTFACT, with six quadrature points, and MAXLOG which again produced the largest errors. When TESTFACT was run with only three quadrature points, it performed worse than NOHARM and the factor analysis approaches, however, it did not perform as bad as MAXLOG.

Bolt and Lall (2003) examined the recovery of item parameters using the Markov chain Monte Carlo estimation and specifically the Metropolis-Hasting algorithm. In their study, they used two sample sizes, 1,000 and 3,000, three correlations, .0,.3, and .6 and two test lengths, 25 and 50 items. The multidimensional two parameter logistic model was employed and they compared the recovery of the item parameters under each condition. The results of the study showed that the item parameter recovery was generally unchanged by the simulation conditions. Tate (2003) also conducted a simulation study to compare parametric and nonparametric methods including Mplus, NOHARM, and TESTFACT for the recovery of dimensionality and item parameters. Tate made use of both unidimensional and multidimensional models, a variety of interdimensional correlations, and different test models including 1PL, 2PL and 3PL models. The overall results of Tate's study showed that Mplus performed well for the 1PL and 2PL models and poorly in the 3PL models. Further results showed that NOHARM and TESTFACT performed positively under the simulation conditions.

Finch (2010 and 2011) expanded on the work done by Bolt and Lall (2003) and Tate (2003) and incorporated more simulation conditions and comparison techniques. Finch (2010) used the unweighted least squares (ULS) estimation technique and the NOHARM software as well as the Robust Weighted Least Squares (RWLS) estimation and the Mplus 3 software to recovery item parameters for a two-dimensional simple structure tests with a variety of test lengths, examinee counts, and interdimensional correlations. Finch (2011) also used the unweighted least squares estimation technique and NOHARM under a 30 item two-dimensional test structure with some items loading onto only one dimension and some items loading onto both dimensions. Along with the change in the test model, Finch (2011) also varied the number of examinees and the correlation between dimensions. In his 2010 study, Finch found that the discrimination and difficulty estimates for ULS and RWLS were not impacted by the change in correlations. Overall, Finch (2010) saw that Mplus 3 produced slightly smaller errors than NOHARM for item parameter recovery. Finch (2011) found that as the correlation between the dimensions increased, the bias between the true and recovered discrimination and difficulty parameters also increased. Finch (2011) also found that items that loaded onto both dimensions produced a larger item parameter bias that those that loaded onto only one dimensions.

Han and Paek (2014) conducted a study that compared IRTPRO 2.1 with BA, MHRM and ADQ estimation techniques, Mplus 7 with the ML and Monte Carlo (MC) integration technique, FlexMIRT with BA and MHRM estimation, and EQSIRT with marginal maximum likelihood (MML), Monte Carlo Expectation-Maximization (MCEM) and the Markov chain Monte Carlo (MCMC) estimation technique. Han and Paek used a compensatory two parameter logistic model with three or four dimensions, varying model complexity, 30 items, and 3,000 examinees. The results of the study showed that the correlations between the true item parameters and the recovered item parameters were high (0.8 and above) for both the item discrimination and the item difficulty. The complexity of the model did not impact the overall recovery of the item parameters.

Chalmers and Flora (2014) investigated both compensatory and noncompensatory models using the MHRM estimation technique. In their study, they used both two- and three-dimensional noncompensatory and compensatory IRT models, different sample sizes (n = 1,000, 2,500, 4,000), a variety of unidimensional items (5,10, 15), and varied the correlation between dimensions (.0, .2, .4, .6, .8). Similar to Bolt and Lall (2003), they found no large changes in the recovery of compensatory item parameters and also found poor recovery for the three dimensional noncompensaroty models especially with small sample sizes and large correlations between dimensions. Babock (2011) investigated the performance of Metropolis- Hasting with Gibbs algorithm in the estimation of a noncompensatory two parameter multidimensional model. In his study, Babcock modified the correlation between dimensions, the sample size, and the number of unidimensional items per dimension. Babcock found that 4,000 examinees was desirable and that higher correlations between dimensions produces poorly estimated *a* parameters.

Previous studies have also investigated the recovery of ability estimates. In 2005, de la Torre and Patz examined the effect of the number of abilities (two and five), number of items (10, 30 and 50), and the degree of correlation between abilities (.0, .4, .7 and .9). All factors were crossed for a total of 24 conditions. The model was a multidimensional 3PL compensatory model with only between item dimensionality. The study demonstrated that as the correlations between abilities, number of items, and number of abilities increased, so did the recovery correlations. The study also showed that with at least 30 items and uncorrelated abilities, the underlying abilities can be estimated reliably. The largest recovery correlations were obtained when there were five dimensions and the correlation between the dimensions was .9.

In another study, de la Torre (2009) used the compensatory multidimensional 3PL model where each item loaded onto only one dimension to investigate the recovery of ability estimates singularly as well as with other ancillary variables such as examinee demographics. Simulation conditions included 10 or 20 items, two to five dimensions, and correlations of .5 and .9 between dimensions. A joint maximum likelihood estimation was used to recover both item parameters and examinee abilities. The simulation produced recovery correlations that were 0.85 and 0.91 for the 10 and 20 item test respectively. Additional results show that the increase in the correlation between dimensions also increased the recovery correlations. Overall, de la Torre found that the inclusion of ancillary variables improved the recovery of ability estimates.

In the aforementioned studies, only models with between item dimensionality were used. Segall (2001) and Wang, Chen and Cheng (2004) used test models that included between and within dimensionality. Segall (2001) used real and simulated ASVAB data with 105 item test which included four dimensions with each item loading onto three of the four dimensions. Segall used the IFACT parameter estimation procedure which is an extension of the MCMC method and Bayesian estimation (Segall, 1998). Segall found that using MIRT almost doubled the reliability results when compared to a simple number correct calculation. Wang, Chen and Cheng (2004) compared the MIRT recovery with the UIRT ability recovery on a 50 item multiple choice test and a 40 item 4 point Likert scale personality inventory. The study utilized Bock Aitkin maximum-likelihood estimation using both ACER ConQuest (Wu, Adams, & Wilson 1998) and SAS MLMIXED (SAS Institute, 1999) procedure. Wang, Chen and Cheng found that multidimensional ability recovery improved with the increased number of dimensions, higher correlations between dimensions, and more items per dimension. Furthermore, they found that dimension with four to nine items produced unreliable ability estimates.

Where previous studies have mostly focused on either item parameter recovery or ability estimates, the current study investigates the recovery of both item parameters and ability estimates. Additionally, many of the simulation conditions, such as correlations between dimensions, differing estimation techniques and a variety of test models are also used in the current study which allows for a more direct comparison of the results. Furthermore, where previous research has been based on two or three dimensional models with only between item dimensionality, the current study goes a bit further and investigates both simple and complex three dimensional test designs.

3 METHODOLOGY

The purpose of this study is to compare the accuracy of item discrimination parameter estimation, item difficulty parameter estimation, and examinee ability estimation for multidimensional item response models using both Mplus 7 and IRTPRO 3. Both Mplus 7 and IRTPRO 3 are used often in unidimensional IRT modeling; nonetheless, there is not a great deal of work comparing their use in the multidimensional setting. A series of simulations was used to evaluate and compare the ability of Mplus 7 and IRTPRO 3 to recover MIRT item parameters and examinee ability. While there have been numerous studies focused on the comparison of MIRT estimation techniques, there is a small body of work that looks at the IRTPRO 3 approach due to the relative newness of the program. This study extends on the body of research by investigating a variety of simulation conditions and includes a comparison of examinee ability estimation.

Methods

The study was conducted in two parts. The first part of the study evaluated and compared the ability of Mplus 7 and IRTPRO 3 to recover the true MIRT item parameters. The second part of the study evaluated and compared the ability of Mplus 7 and IRTPRO 3 to recover the true examinee ability. This study was conducted with the practitioner in mind and was set up in such a way to mimic real world conditions.

Item parameter estimation. Item parameters for a compensatory two parameter logistic MIRT model were used. There are a variety of models that could have been chosen, however, the compensatory two parameter model was chosen due to its use in other studies (Bolt & Lall 2003, Finch 2010; Finch 2011; Knol & Berger, 1991; Tate, 2003) The item discrimination parameter values are similar to a simulation run by Reckase (2009) in which he generated item parameters for a three-dimension test and used them to investigate the model estimation procedures for four different MIRT modeling software. Since the recovery of item parameter estimates were of interest, the first step in the simulation was to establish a sample size that was large enough to produce relatively stable item parameter estimates. Previous simulations studies have used sample sizes as small as 100 and as large as 4,000; though these studies were conducted with multiple repetitions (Babcook, 2011; Bolt, 2003; Carrol, Williams, & Levine 2007; Finch, 2011; Han & Paek, 2014; Kahrama, 2013; Wang & Nydick, 2015; Wen-Chung, 2004). The focus of this study was on the use of the software programs in the real world. Practitioners are often only given one data set and therefore, this simulation study was run using one main run and two verification replications.

A simple structure, 24 item, multidimensional model was used to establish a stable sample size. In this model, one-third of the items load onto the first factor, one-third of the items load onto the second factor and the final third of the items load onto the third factor. The structure of the test can be likened to the structure of a general math test where some items are related to algebra, other to geometry, and yet others to calculus. This model can be seen in Figure 5. The item parameters for Model 1are shown in Table 1.

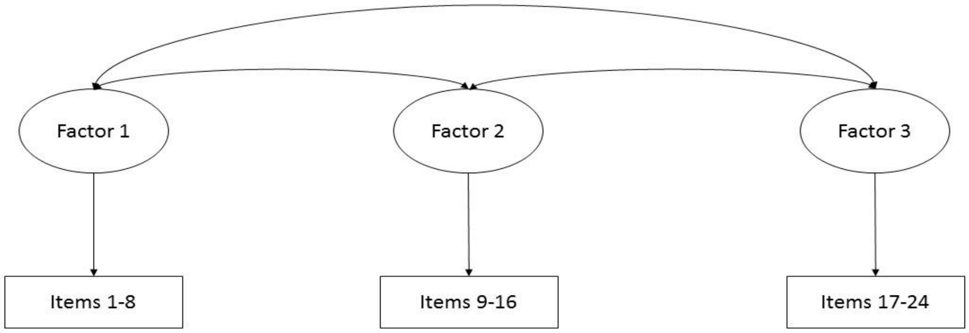

Figure 5. Between-item multidimensional IRT model

Table 1.

Item Parameters Between-item multidimensional IRT model Model 1

Item number	a_1	a_2	a_3	d
1	0.04	0.00	0.00	0.08
2	0.02	0.00	0.00	-0.09
3	0.15	0.00	0.00	-0.29
4	0.09	0.00	0.00	-0.29
5	0.16	0.00	0.00	-0.22
6	0.16	0.00	0.00	-0.31
7	0.20	0.00	0.00	-0.63
8	0.12	0.00	0.00	0.40
9	0.00	0.04	0.00	0.04
10	0.00	0.05	0.00	0.05
11	0.00	0.14	0.00	0.45
12	0.00	0.02	0.00	-0.14
13	0.00	0.04	0.00	0.23
14	0.00	0.03	0.00	-1.02
15	0.00	0.02	0.00	0.23
16	0.00	0.00	0.00	-0.16
17	0.00	0.00	0.05	-0.16
18	0.00	0.00	0.12	0.31
19	0.00	0.00	0.06	0.71
20	0.00	0.00	0.05	0.03
21	0.00	0.00	0.06	0.38
22	0.00	0.00	0.11	-0.12
23	0.00	0.00	0.15	-0.17
24	0.00	0.00	0.01	-1.07

As with other MIRT simulation studies, examinee ability estimates were taken from a multivariate normal distribution with a mean of zero and a standard deviation of one (Bolt, 2003; Carrol, Willimas, & Levine 2007; Finch, 2010; Finch, 2011; Finch & Habing, 2005; Kahraman, 2013; Wang & Nydick, 2015; Wen-Chung, 2004). Using the given item parameters and an uncorrelated multivariate normal theta values, simulated examinee responses were produced from SAS such that for every examinee, $x_{i,}$ the examinee response was calculated using Equation 20 where u_{vi} was taken from the random uniform distribution.

Equation 20 Examinee Response

$$x_{vi} = \begin{cases} 1\ if\ p_i(\theta_v) > u_{v1} \\ 0, if\ p_i(\theta_v)\ \leq u_{v1} \end{cases} \tag{20}$$

The WLSMV and Bayes estimation for Mplus 7 and the BA approach with expectation maximizing algorithm, ADQ and MHRM approach for IRTPRO 3 were run to estimate the item parameters. The sample size varied from 1,000 to 10,000 in increments of 1,000.

To establish a sufficient sample size, the multidimensional discrimination (MDISC) and multidimensional difficulty parameters were compared using the root mean square error (RMSE). The equation for RMSE is shown below where, $\hat{x}_i$ is the average parameter value obtained from all of the replications, x_i is the true value and *n* is the number of test items.

Equation 21 Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{21}$$

After establishing a sufficient sample size, a simulation with that sample size, a simple three-dimensional structure with zero correlations between each factor, and 24 items with 8 items loading on each dimension was used to identify the top estimation technique for each program. These estimation techniques were then used in the rest of the simulation conditions which included three models and three correlations.

The first model is identical to the model used to identify an adequate sample size and the top performing estimation technique for each program. The second model used in this study is the model used by Kim, et al. (2013) and is shown in Figure 6. As stated in their study, the within-item multidimensional model is suitable when an item measures two or more abilities

(Kim et al., 2013). The model used in this study not only contained within-item dimensionality, it also contained some items that only load onto one factor, between-item dimensionality. The inclusion of both within-item and between-item dimensionality made this model a between- and within-item MIRT model and is named as such. Half of the items loaded onto a single factor and the other half of the items loaded onto two factors. A real-world example of this test structure is a math test where there are questions that assess computation, algebra, and calculus separately. Some questions assess both algebra and calculus, some assess both computation and calculus, and others assess both computation and algebra.

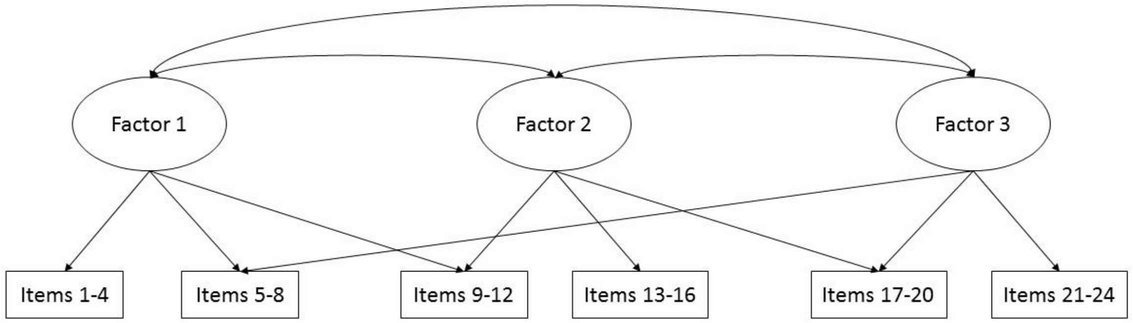

Figure 6. Between and within-item multidimensional IRT model

The final model is another within- and between-item multidimensional model. The final model in this study, shown in Figure 7, is analogous to one model used by Wang and Nydick (2015) in which half of the items loaded exclusively onto one of the three factors, one-fourth of the items load onto two factors and one-fourth load onto all three factors. Wang and Nydick randomly selected the items that would load onto two or three factors. For this study, the items that load onto two or three abilities were not randomly selected, but were identified before the simulation was run. The factor combination varied for items that assessed two factors. One third of the two factors items load onto factors one and two, one-third of the two factors items load onto factors two and three, and one-third of the two factors items load onto factors one and three.

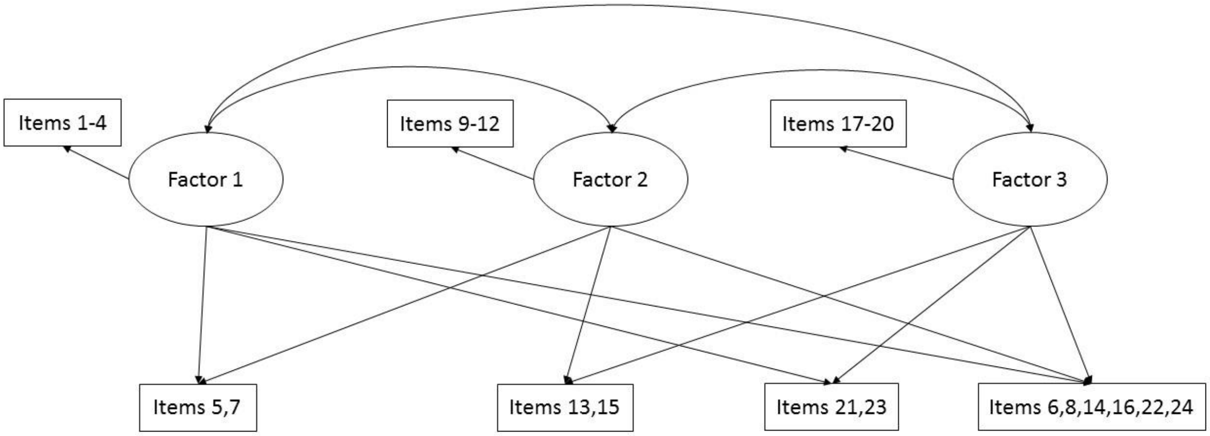

Figure 7. Complex between and within-item multidimensional IRT model

The correlations between the factors range from .0 to .9 in increments of three analogous to the simulation done by Bo and Stone (2008). This was done in order to simulate data in which factors have varying levels of correlations from relatively small to large.

All models were simulated with all conditions for a total of nine conditions. As stated previously, three replications were run in order to more accurately represent the conditions that a practitioner would encounter. Having three replications allows for a comparison of estimation approaches for each simulation condition and it also allows for the observation of fluctuation over replications. The results of each replication were compared against each other as well as compared against the true values. RMSE was again used to make the comparison. All item parameters for the models are shown in Appendix A. The defaults of the program were run when possible to understand how each program might be used in a real-world setting.

Examinee ability estimation. The second part of the study focused on examinee ability estimation. The investigation of each estimation technique to recover examinee ability estimates was set up in a similar fashion to the first part of the study. First, a sufficient number of items per dimension was established. This was done using the same simple three-dimensional structure with no correlation between factors and one-third of the items loading exclusively onto one of the three dimensions. The number of items per dimension varied from eight to 24 in increments of two. Again, examinee ability estimates were taken from a multivariate normal distribution with a mean of zero and a standard deviation of one. SAS was used to produce examinee response patterns, and three replications were run. The sample size used came from the first part of the simulation study. To understand the examinee ability recovery for each estimation technique, Pearson's correlations were used. The ability estimates were compared to the true values for each dimension.

After identifying the appropriate number of items per dimension, another simulation with a simple three-dimensional structure with zero correlations between each factor was used to identify the top preforming estimation technique for each software program. These top two software programs were then exposed to the same nine simulation conditions as in the first part of the study, three different models, three different correlations, and three replications. All item parameters are shown in Appendix A.

4 RESULTS

Before embarking on the multidimensional aspect of the study, the ability of each estimation technique to recover the item parameters and examinee abilities were run in the simplest case to serve as a baseline. The baseline model used for this study was a simple unidimensional 24 item exam with 10,000 examinees. The model serves as a comparison of the results from the more complex multidimensional models. If the estimation techniques did not recover the true parameters well for the baseline model, then that would suggest that the addition on any complexity to the models would decrease the estimations techniques ability to recover true item parameters and examinee ability.

The results show that the discrimination parameters and the difficulty parameters were recovery relatively well. Additionally, the correlations between the true abilities and the recovered abilities were relatively high. The item parameter results of each of the estimation techniques for each of the repetitions are shown in Figures 8 and 9, and the results of the correlations between examinee abilities and recovered abilities for each repetitions and technique are shown in Table 2.

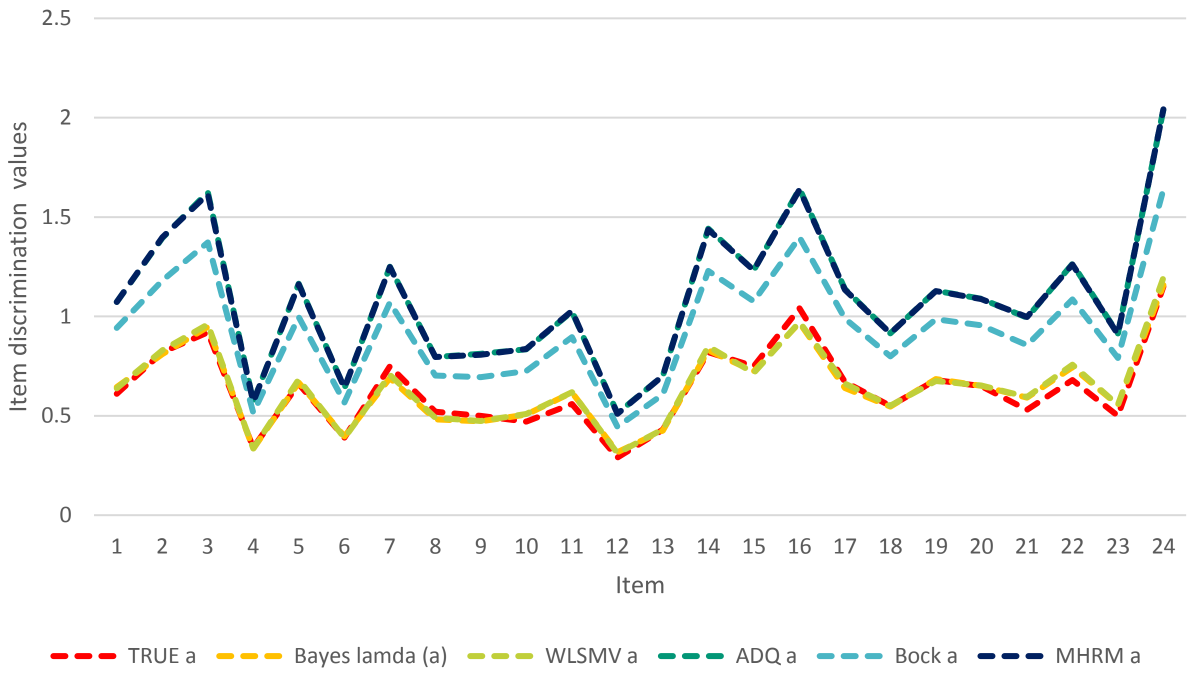

Figure 8. Unidimensional discrimination parameter recovery

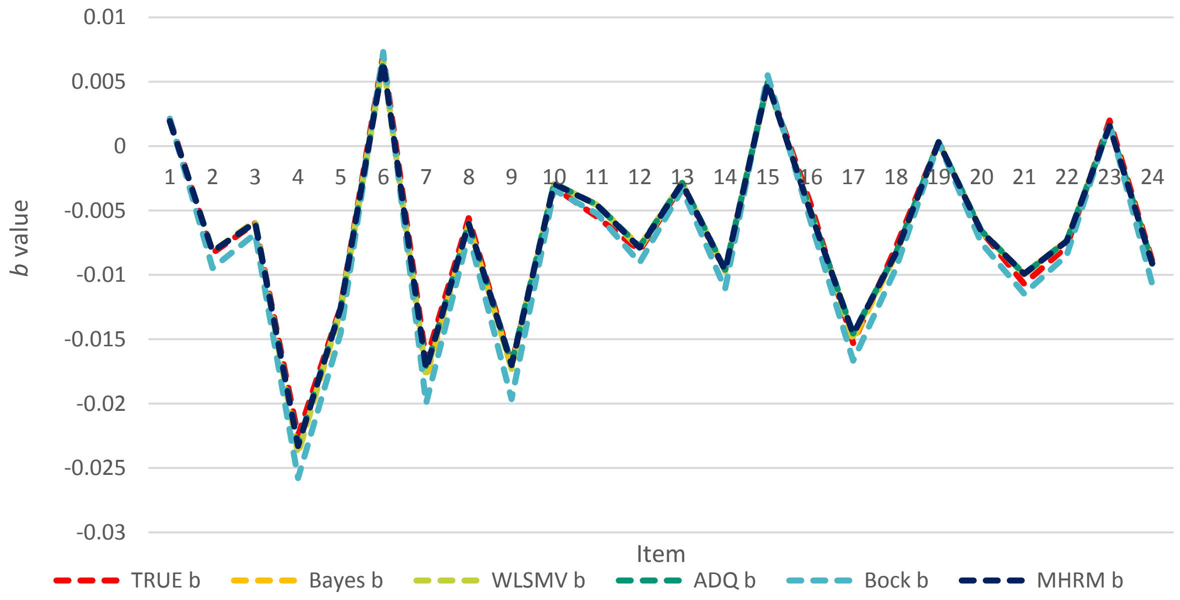

Figure 9. Unidimensional difficulty parameter recovery

Table 2.

Correlations between true abilities scores and recovered ability estimates

	True Ability	BA	ADQ	MHRM	WLSMV	Bayes Mean	Bayes Median
True ability	1						
BA	0.887**	1					
ADQ	0.903**	0.980**	1				
MHRM	0.795**	0.861**	0.878**	1			
WLSMV	0.903**	0.979**	1.000**	0.878**	1		
Bayes Mean	0.894**	0.969**	0.989**	0.868**	0.989**	1	
Bayes Median	0.890**	0.964**	0.984**	0.863**	0.985**	0.996**	1

* *p < 0.001

Item Parameter recovery

As stated in the methods sections, the first part of the study investigated the ability of each estimation technique to recover the true item parameters. Before true parameters could be recovered, it was necessary to identify an adequate sample size. To find the sample size needed, a simple three dimensional 24-item test was used and sample size varied from 1,000 to 10,000. The following describes the results of the sample size exploration

Adequate sample size. The defaults for the BA estimation technique in IRTPRO 3 was 49 quadrature points and for the ADQ the default was nine quadrature points. When theses defaults were used, both estimation techniques did not converge. Schilling and Bock (2005) express that the number of quadrature points must decrease when the number of dimensions increase. Because of this, when the ADQ estimation was rerun, 2 points per dimension was used for a total of six quadrature points. Similarly, three quadrature points per dimension were used with the BA estimation technique for nine quadrature points.

Although the intent was to run only three replications for each on the sample sizes, the WLSMV estimation technique failed to converge for any of the repetitions with the 3,000, 5,000, and 7,000 examinee samples and for the 1,000, 2,000 4,000, 6,000, 8,000, and 10,000 samples, only one of the three replications converged. In order to produce a sufficient number of replications, examinee responses were simulated until there were at least two converging replications per sample size.

After averaging the RMSE values together for each of the replications, the change in RMSE became relatively small after the 8,000 examinee sample size. For this reason, a sample size of 8,000 was determined to be sufficient. The results depicted that each estimation technique varied greatly with respect to the sample size. The multidimensional difficulty error values consistently decreased as the sample size increased. The same is true for the MDISC parameter, but, the *d* parameter did not experience as much of a decrease as the MDISC parameter. The results are provided in Figures 10 and 11.

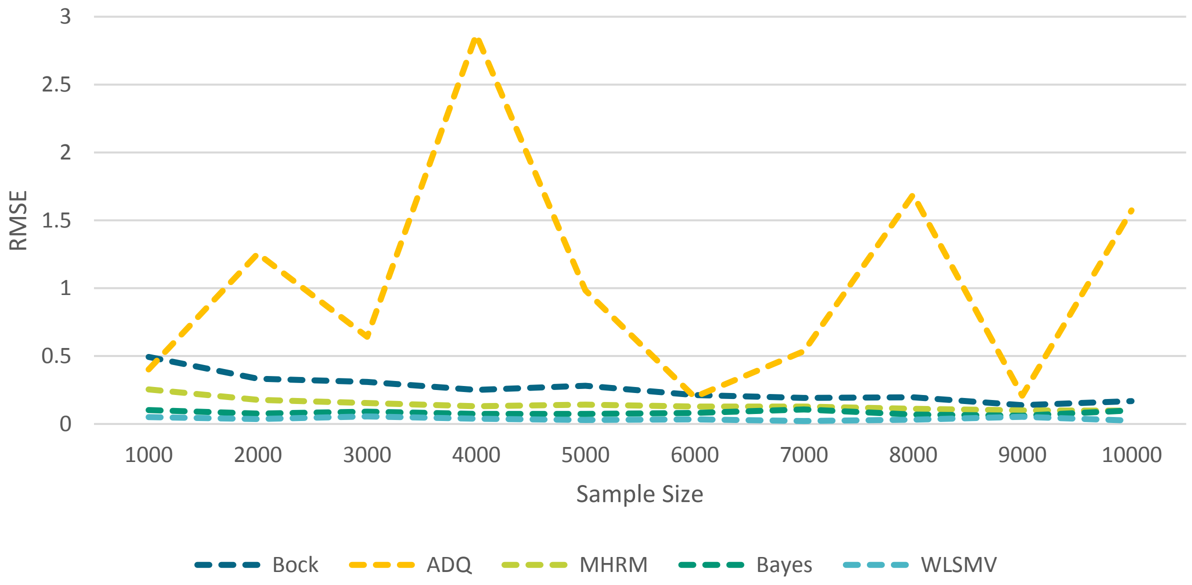

Figure 10. MDISC recovery for adequate sample size

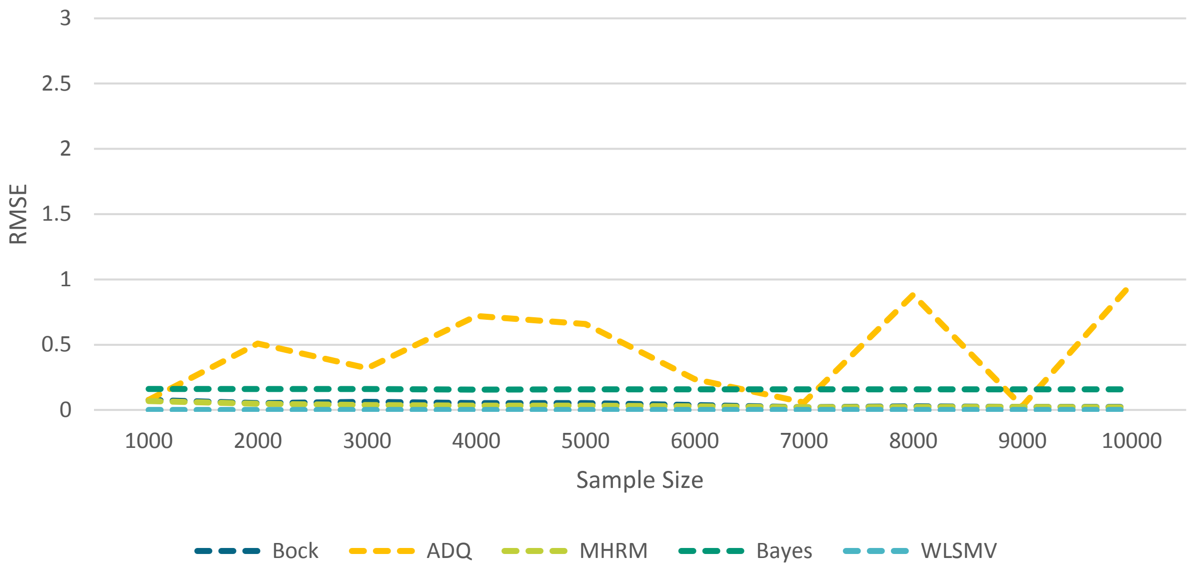

Figure 11. Difficulty recovery for adequate sample size

After a sufficient sample size was established, the top estimation techniques for each software program were selected. Although the average RMSE of the WLSMV estimation technique was consistently less than that of the Bayes estimation technique, not having any of the replication converge was far too risky. Because of the WLSMV convergence issues, the Bayes estimation technique was selected from Mplus 7.

With the exception of the ADQ estimation technique, the average RMSE for the IRTPRO 3 estimation techniques were relatively stable. Each of the replications converged and produced estimates. The RMSE for the ADQ technique varied quite a bit. The literature suggested that the ADQ technique works best in complex models, and this might be one reason the values had the greatest variance of all estimation techniques across the software. Using a different number of quadrature points might have produced a different result; however, the investigation of quadrature points was not the focus of this study. The results demonstrated that the BA and the MHRM estimation techniques produced similar, relatively small, RMSE values.

Although BA is the default estimation technique in IRTPRO 3, MHRM was chosen because a previous pilot study showed that IRTPRO 3 consistently produced an error when the BA estimation technique was used with multiple dimensions per item. In the end, a sample size of 8,000 examinees with both the Bayes estimation technique from Mplus 7 and the MHRM estimation technique from IRTPRO 3 was selected for use the recovering item parameters in the remaining conditions.

Conditions 1 through 4. Model 1 was used for the first four conditions. The only variable that changed between conditions was the correlation between factors. The results of the first four conditions portrayed the Bayes estimation technique with a lower RMSE than the MHRM technique when recovering the MDISC parameter. Conversely, the results exhibited a lower average *d* RMSE for the MHRM technique than the Bayes estimation technique. The first replication yielded the MDISC RMSE with the biggest difference between the two estimation techniques and the second replication yielded the biggest difference in the *d* RMSE for the two estimation techniques.

The results for Condition 2 were similar to that of Condition 1. Again, the RMSE MDISC average recovery was the smallest with the Bayes estimation technique and the RMSE *d* average recovery was the smallest with MHRM technique. The overall RMSE was very similar to the ones seen in the first condition. Although the increase from a zero correlation to a .3 correlation was reasonably large, the results illustrated that the increase in correlations did not have a large impact on the parameter recovery. It should be noted, that some researchers place both correlations .0 and .3 in the small category (Hinkle, Wiersma & Jurs, 2003) and therefore might not consider a change in correlation from .0 to .3 which would be reasonably large.

Condition 3 moves the correlation into the medium category with a correlation of .6 (Hinkle, Wiersma & Jurs 2003). Although the correlation increased, each software program performed in a similar matter to as in the first two conditions. Again, Bayes produced closer MDISC estimates and MHRM produced closer *d* parameter estimates. Of the conditions based on Model 1, Condition 3 yielded the highest average RMSE values for both software program. The results also depicted that the RMSE *d* recovery values were somewhat stable across the three replications.

Condition 4 had the highest correlations out of the four condition at .9. A correlation of this size is considered large (Hinkle, Wiersma & Jurs 2003). With such a large correlation between dimensions, the model is almost unidimensional. As with Condition 3, the *d* RMSE recovery was moderately stable and in Condition 4, the RMSE for MDISC was also relatively stable. The stability of the estimations are illustrated in Figures 12 and the overall results of the four conditions and averages for Model 1 are given in Table 3.

Table 3.

Average MDSIC RMSE and *d* RMSE by condition

	Bayes		MHRM	
	Average MDSIC RMSE	Average *d* RMSE	Average MDSIC RMSE	Average *d* RMSE
Condition 1	0.073	0.160	0.214	0.052
Condition 2	0.076	0.160	0.108	0.023
Condition 3	0.064	0.156	0.094	0.024
Condition 4	0.060	0.160	0.108	0.024

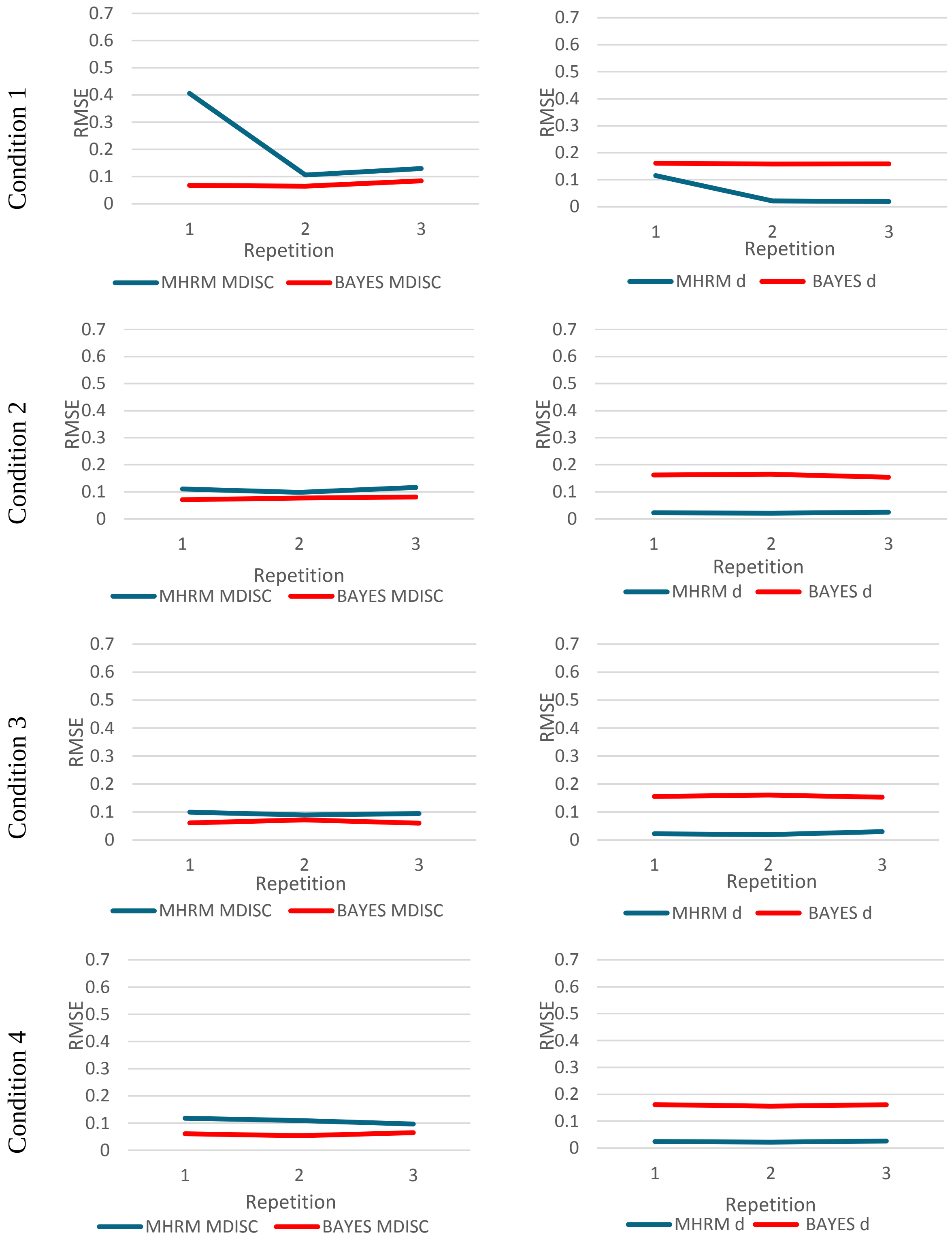

Figure 12. MDISC and *d* parameter recovery for conditions one through four

Conditions 5 through 8. Model 2 was used for the next set of conditions. Again, the only variable that changed between conditions was the correlation between dimensions. Model 2 included within-item and between-item dimensionality. The correlation between dimensions was zero for Condition 5. The average RMSE for the MDISC recovery was the smallest with the Bayes estimation technique and for the *d* recovery and the average RMSE was the smallest with the MHRM estimation technique. The results also showed that the *d* RMSE recovery values were comparatively close for replication one.

Conditions 6 and 7 produced similar results as seen in Condition 5. For these conditions, Bayes estimation out preformed MHRM estimation for MDISC recovery while MHRM estimation outperformed Bayes estimation for *d* parameter recovery. The results showed a larger discrepancy between estimation techniques in recovering the MDISC than seen in the first four conditions. Additionally, the discrepancies in the *d* parameter recovery varied more under Model 2 than it did in Model 1.

The correlations for Condition 8 were all set at .9, again making them similar to the unidimensional model. Just as in with the previous conditions, in Condition 8 the Bayes estimation technique produced the smallest RMSE for the MDISC and MHRM produced the smallest RMSE for the *d* parameter recovery. As in the first four conditions, the difference between the *d* recovery for each estimation technique is rather large for each replication. Additionally, the *d* RMSE results are reasonably stable for each estimation techniques. The results for conditions five through eight are shown in Figure 14 and Table 4.

Table 4

. Average MDSIC RMSE and *d* RMSE by condition

	Bayes		MHRM	
	Average MDSIC RMSE	Average *d* RMSE	Average MDSIC RMSE	Average *d* RMSE
Condition 5	0.146	0.153	0.438	0.053
Condition 6	0.144	0.156	0.435	0.022
Condition 7	0.162	0.330	0.446	0.025
Condition 8	0.166	0.150	0.444	0.022

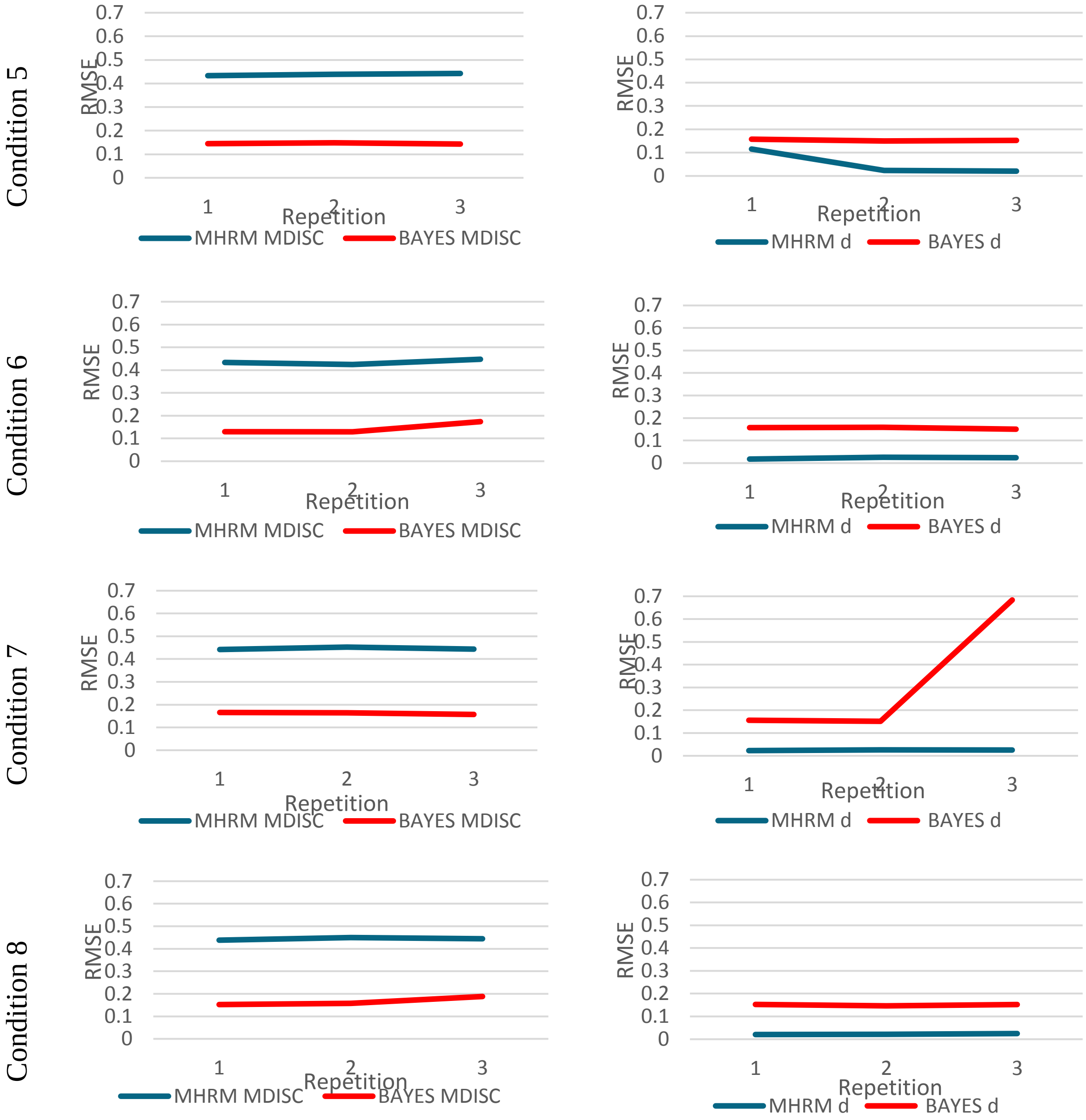

Figure 13. MDISC and *d* parameter recovery for conditions five though eight

Conditions 9 through 12. Model 3 was the basis for the final set of conditions. Just as in the previous conditions, the only change between conditions nine through twelve was the increased correlations between dimensions. Model 3 also included both within- and between-item dimensionality and the most complexity of the three models. Overall, the results of Condition 9 were similar to the results of Conditions 1 through 8. In Condition 9, the Bayes estimation technique produced the smallest RMSE for the MDSIC recovery and the MHRM estimation technique produced the smallest RMSE for the *d* recovery.

The results from Conditions 10 and 12 did not change much from the results of Condition 9 even though the correlations between the two dimensions increased. Again, the Bayes estimation technique proved best for the MDISC recovery and the MHRM estimation technique proved best for the *d* parameter recovery. Condition 12 was a little different. In general, the results were the same with Bayes being the best with MDISC recovery and MHRM being the best for the *d* recovery, however, the difference between the average MDISC recovery to each estimation technique was the largest in this condition. The graphs for conditions nine through twelve are shown in Figure 14 and Table 5 shows the overall results of all four conditions and averages for Model 3.

Table 5.

Average MDSIC RMSE and *d* RMSE by condition

	Bayes		MHRM	
	Average MDSIC RMSE	Average *d* RMSE	Average MDSIC RMSE	Average *d* RMSE
Condition 9	0.317	0.315	0.491	0.053
Condition 10	0.256	0.142	0.502	0.022
Condition 11	0.249	0.148	0.545	0.025
Condition 12	0.297	0.146	0.594	0.022

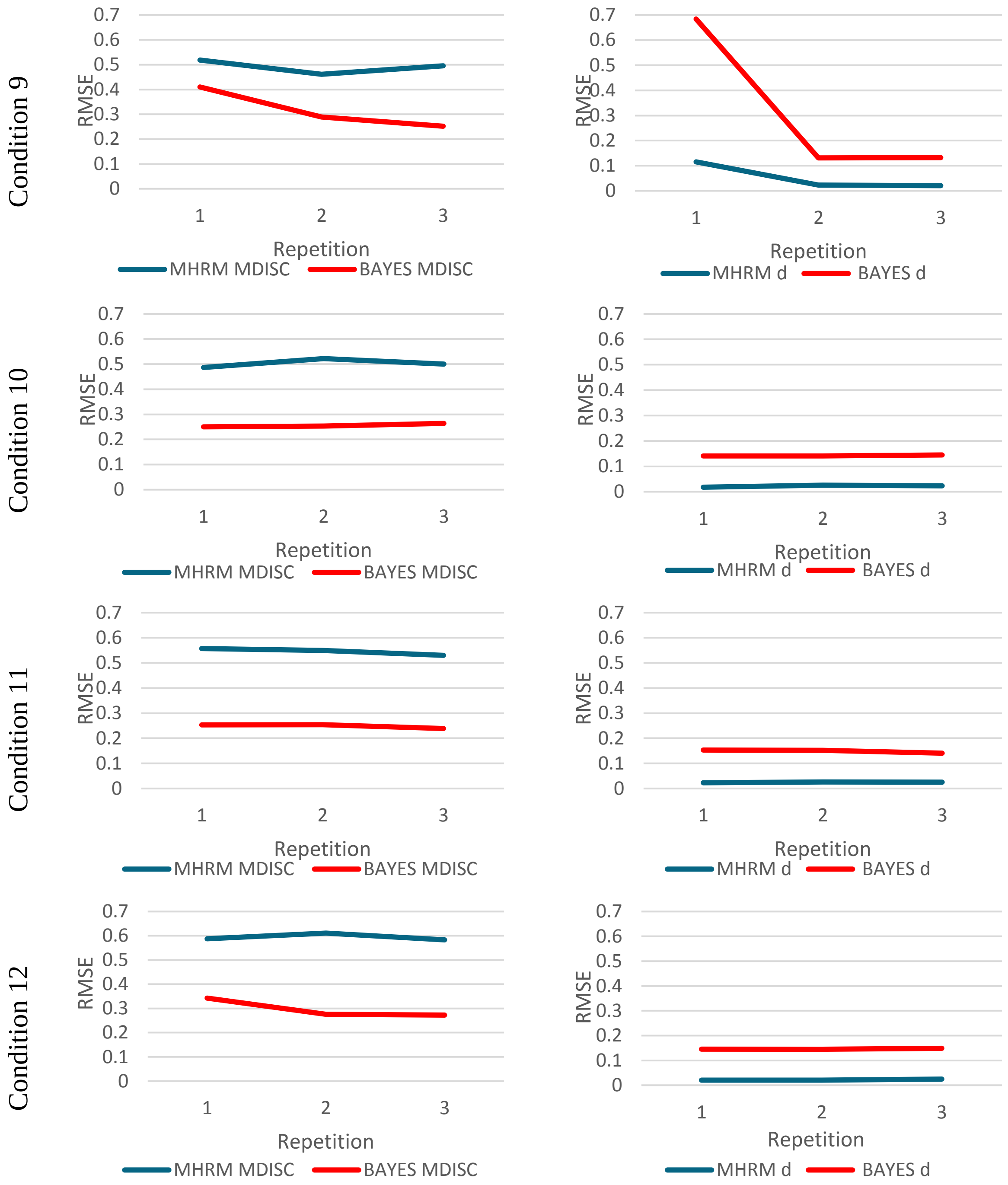

Figure 14. MDISC and *d* parameter recovery for conditions nine through twelve

Examinee ability estimation

Before commencing on the comparison of the examinee ability recovery, it was important to identify a sufficient number of items per dimension. A simple three-dimensional test with 10 to 24 items per dimensions was used to identify the optimal number of items needed. As stated previously, this range was chosen based on the results from the first part of the study where between eight and fourteen items loaded onto one dimension. The IRTPRO 3 ADQ estimation technique did not produce ability estimates and, therefore, the expected a posteriori (EAP) method was used to obtain ability estimates.

Items per dimension. As with the first part of the study, three replications were used for each of the item sizes. The initial results yielded correlations between the true ability and the recovered ability as low as .021. The low recovery correlations were only seen in the recovery of the third dimension. This caused the average correlation values for each estimation technique to be skewed because there were only three replications and one was an outlier. The initial recovery correlation results are shown in Figure 15 and Tables 6 through 8.

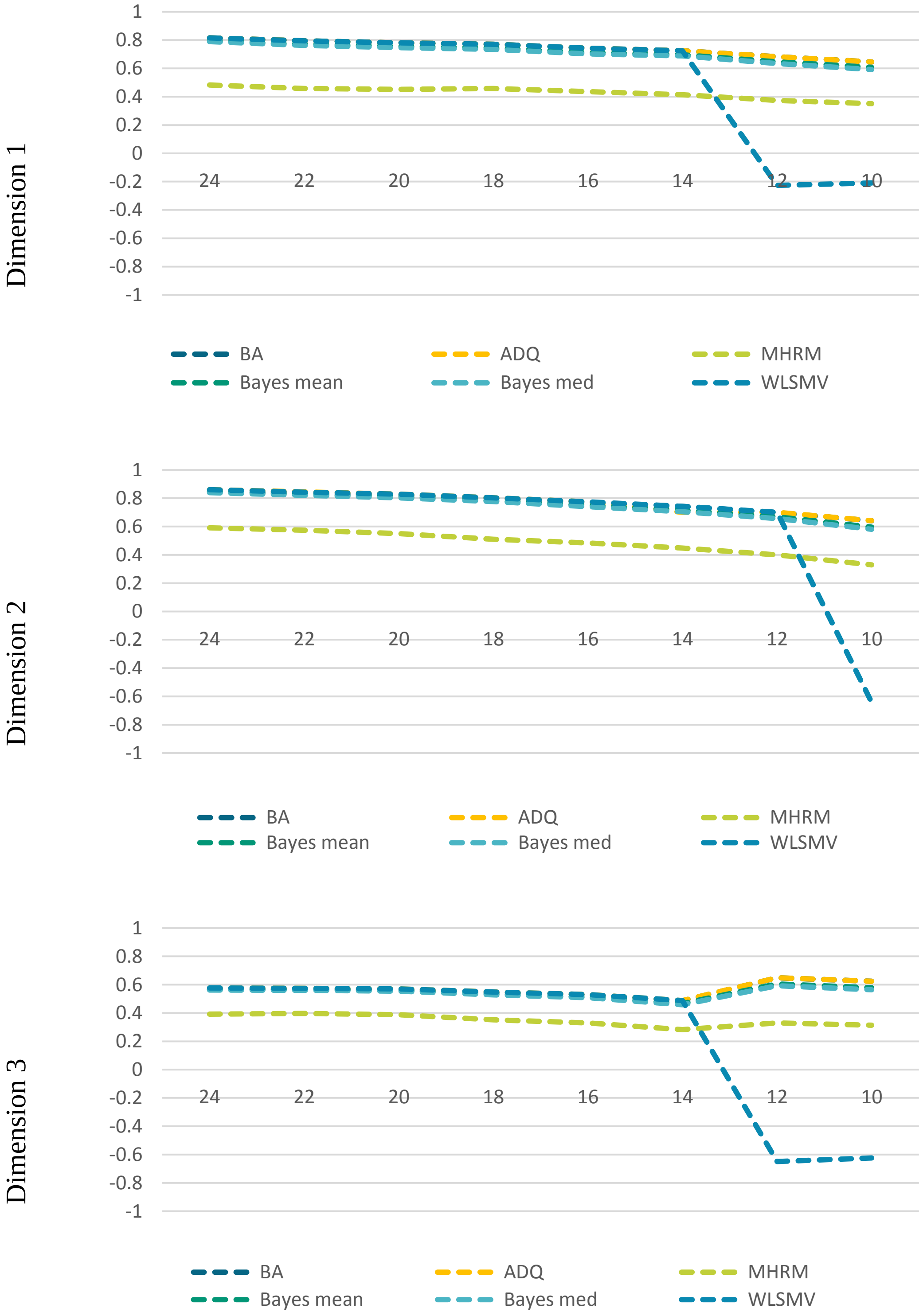

Figure 15. Examinee ability recovery by dimensions and items per dimension

Table 6.

Dimension 1 correlations between examinee ability estimates and true ability

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm M SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.815	0.002	0.815	0.002	0.489	0.012	0.796	0.003	0.789	0.003	0.815	0.002
22	0.795	0.002	0.795	0.002	0.461	0.006	0.772	0.001	0.763	0.002	0.795	0.001
20	0.779	0.003	0.779	0.003	0.45	0.014	0.756	0.003	0.746	0.003	0.779	0.003
18	0.769	0.003	0.769	0.003	0.451	0.013	0.743	0.005	0.733	0.006	0.769	0.003
16	0.741	0.005	0.741	0.005	0.432	0.008	0.711	0.005	0.701	0.005	0.740	0.005
14	0.725	0.002	0.725	0.002	0.414	0.000	0.695	0.003	0.686	0.005	0.725	0.002
12	0.683	0.003	0.683	0.003	0.374	0.009	0.647	0.004	0.635	0.005	-0.226	0.789
10	0.646	0.014	0.646	0.014	0.351	0.022	0.608	0.016	0.592	0.015	-0.210	0.748

Table 7.

Dimension 2 correlations between examinee ability estimates and true ability

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm M SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.846	0.006	0.857	0.006	0.589	0.009	0.842	0.006	0.836	0.007	0.857	0.006
22	0.829	0.010	0.838	0.010	0.570	0.008	0.822	0.010	0.816	0.010	0.838	0.010
20	0.817	0.007	0.825	0.006	0.549	0.003	0.808	0.005	0.801	0.004	0.825	0.006
18	0.795	0.006	0.800	0.006	0.511	0.003	0.779	0.006	0.772	0.008	0.800	0.006
16	0.768	0.004	0.772	0.004	0.479	0.010	0.749	0.005	0.739	0.005	0.772	0.004
14	0.740	0.005	0.700	0.005	0.448	0.013	0.714	0.005	0.704	0.004	0.742	0.005
12	0.699	0.004	0.700	0.004	0.400	0.009	0.668	0.007	0.656	0.006	0.700	0.004
10	0.641	0.005	0.642	0.005	0.329	0.015	0.596	0.008	0.581	0.011	-0.642	0.005

Table 8.

Dimension 3correlations between examinee ability estimates and true ability

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm M SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.569	0.476	0.577	0.480	0.392	0.330	0.567	0.472	0.563	0.469	0.577	0.480
22	0.569	0.462	0.575	0.467	0.397	0.313	0.564	0.458	0.559	0.454	0.575	0.467
20	0.564	0.450	0.570	0.455	0.387	0.314	0.559	0.445	0.554	0.442	0.570	0.455
18	0.543	0.449	0.548	0.451	0.352	0.305	0.534	0.439	0.528	0.435	0.547	0.450
16	0.528	0.434	0.531	0.435	0.330	0.271	0.513	0.424	0.508	0.419	0.530	0.435
14	0.487	0.404	0.488	0.404	0.283	0.232	0.467	0.385	0.459	0.380	0.488	0.404
12	0.649	0.005	0.649	0.005	0.331	0.004	0.606	0.002	0.593	0.003	-0.649	0.005
10	0.624	0.008	0.624	0.007	0.314	0.006	0.578	0.007	0.564	0.007	-0.624	0.007

Because of the drastic differences in recovery of the third dimension, the mean correlations are pulled down and the standard deviations are quite large. A modified *z*-score equation, as seen in Equation 17, where M_i is the modified z-score, x_i is the data point, and $\tilde{x}$ is the median, was used to identify outliers (Iglewicz and Hoaglin, 1993). In order to combat the skew in the results, repetitions were run until there were at least two replications that produced estimates that were not considered outliers by using the modified *z*-score technique. The estimates from the two replications were then averaged together. For most item loading sample sizes, only one other replication needed to be run, however in the case of 22 items per dimension, two extra replications were needed in order to produce recovery correlations in the acceptable range. Additionally, seven replications had to be run for the 16-items per dimension size. The updated means and standard deviations are displayed in Figure 16 and Tables 9 through 11. The updated results provided means that were not affected by outliers and the standard deviations were smaller than seen previously in the third dimension.

Equation 22 Modified *z* –score

$$M_i = \frac{0.6745(x_i - \widetilde{x)}}{median(|x_i - \tilde{x}|)} \qquad (22)$$

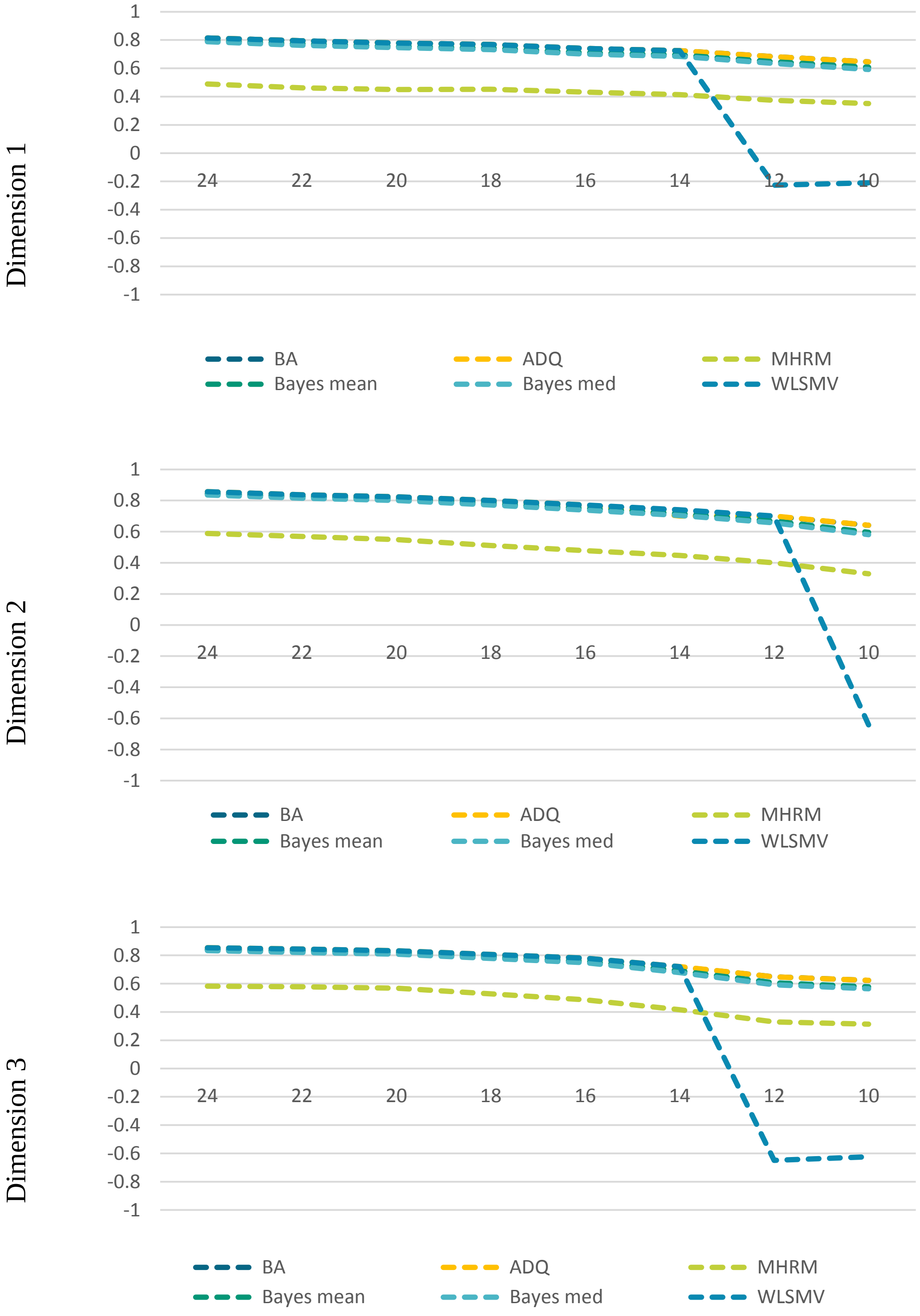

Figure 16. Examinee ability recovery by dimensions and items per dimension

Table 9.

Dimension 1 correlations between examinee ability estimates and true ability with no outliers

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.816	0.002	0.816	0.002	0.483	0.012	0.796	0.003	0.790	0.003	0.816	0.002
22	0.795	0.002	0.795	0.002	0.458	0.003	0.771	0.001	0.763	0.002	0.795	0.002
20	0.780	0.004	0.780	0.004	0.451	0.020	0.756	0.004	0.746	0.004	0.780	0.003
18	0.770	0.003	0.770	0.003	0.459	0.002	0.745	0.005	0.736	0.005	0.770	0.003
16	0.743	0.004	0.743	0.004	0.436	0.006	0.712	0.006	0.703	0.006	0.743	0.004
14	0.724	0.001	0.724	0.001	0.414	0.001	0.697	0.001	0.688	0.003	0.724	0.001
12	0.683	0.003	0.683	0.003	0.374	0.009	0.647	0.004	0.635	0.005	-0.226	0.789
10	0.646	0.014	0.646	0.014	0.351	0.022	0.608	0.016	0.592	0.015	-0.210	0.748

Table 10.

Dimension 2 correlations between examinee ability estimates and true ability with no outliers

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.849	0.006	0.861	0.006	0.591	0.009	0.845	0.006	0.840	0.007	0.860	0.006
22	0.835	0.001	0.844	0.002	0.575	0.000	0.828	0.003	0.822	0.003	0.844	0.002
20	0.821	0.003	0.828	0.002	0.550	0.003	0.811	0.000	0.804	0.000	0.828	0.002
18	0.798	0.001	0.804	0.001	0.510	0.004	0.783	0.001	0.776	0.001	0.804	0.001
16	0.771	0.003	0.774	0.004	0.484	0.003	0.751	0.005	0.741	0.004	0.774	0.003
14	0.742	0.004	0.700	0.005	0.447	0.018	0.717	0.005	0.705	0.003	0.744	0.005
12	0.699	0.004	0.700	0.004	0.400	0.009	0.668	0.007	0.656	0.006	0.700	0.004
10	0.849	0.006	0.861	0.006	0.591	0.009	0.845	0.006	0.840	0.007	0.860	0.006

Table 11.

Dimension 3 correlations between examinee ability estimates and true ability with no outliers

Items per dim	BA r	BA SD	ADQ r	ADQ SD	mhrm r	mhrm SD	Bayes mean r	Bayes mean SD	Bayes Median r	Bayes Median SD	wlsmv r	wlsmv SD
24	0.843	0.001	0.854	0.001	0.583	0.008	0.840	0.001	0.834	0.001	0.854	0.001
22	0.835	0.003	0.844	0.003	0.578	0.021	0.828	0.004	0.822	0.004	0.844	0.003
20	0.824	0.001	0.832	0.001	0.569	0.001	0.816	0.001	0.809	0.002	0.833	0.001
18	0.802	0.008	0.808	0.009	0.528	0.013	0.787	0.009	0.779	0.008	0.807	0.009
16	0.778	0.004	0.782	0.003	0.487	0.003	0.758	0.002	0.749	0.001	0.782	0.003
14	0.720	0.001	0.721	0.002	0.416	0.003	0.690	0.003	0.678	0.003	0.721	0.002
12	0.649	0.005	0.649	0.005	0.331	0.004	0.606	0.002	0.593	0.003	-0.649	0.005
10	0.624	0.008	0.624	0.007	0.314	0.006	0.578	0.007	0.564	0.007	-0.624	0.007

As expected, the results illustrated that the correlation between the true and the recovered examinee ability estimations decreased as the number of items that load on a dimension decreased. Although there is no major dip in the graphs, the lack of reliability of the 16-items per dimension caused a red flag and therefore 18-items per dimension was selected as the minimum number of items per dimension. Additionally, 18-items per dimension was selected because two of the original three replications produced estimates within the acceptable range.

For the estimation techniques, the MHRM technique was ruled out immediately due to the comparatively low recovery correlations even though the recovery correlations were relatively stable. The MHRM recovery correlations were four to five-tenths of a point smaller than the other techniques. The remaining two estimation techniques in the IRTPRO 3 software were very close, so much so that the line graphs are indistinguishable. Because the results from the first section of the study were being combined with the second part of the study, the ADQ technique was selected because the BA estimation technique would not run in complex models. Although the ADQ estimation technique varied greatly, it did produce results. The Bayes estimation technique was again selected because it was also used in the first part of the study. The Mplus 7 Bayes estimation technique produced both the mean and median ability estimate for each of the ten imputations. The correlations for this technique were also very stable and relatively high.

Conditions 1 through 4. The first four conditions were based on Model 1 with the correlations between dimensions increasing in each subsequent conditions. In Model 1, 18 unique items loaded onto each dimension with no cross loading. The Bayes median ability recovery consistently produced the smallest correlations between the true abilities and the recovered abilities. The ADQ estimation technique consistently produced the largest correlations between true ability and recovered ability. The difference between the recovery correlations was the smallest in Dimension 3 for all four conditions, indicating that the results for Dimension 3 were the most consistent between estimation techniques. Overall, the difference between ability recovery correlations were relatively small. When compared to other conditions, the recovery correlations decreased as the correlation between dimensions increased.

An examination of each individual dimension showed that the correlations between the true abilities and the recovered abilities were the smallest for Dimension 1 and the largest for Dimension 3. These results were exhibited in all conditions related to Model 1. The results also indicted that that as the correlation between dimensions increased, the correlation between the recovered abilities and the true abilities also increased. The increase was also seen in all four conditions. The graphical and tabular results for the first four models are shown in Figure 17 and Table 12.

Table 12.

Average correlation and standard deviations for each dimension for conditions 1-4

Condition	Dimension	Bayes mean r	Bayes mean SD	Bayes median r	Bayes median SD	ADQ r	ADQ SD
1	1	0.748	0.004	0.739	0.002	0.775	0.004
	2	0.777	0.004	0.769	0.005	0.798	0.004
	3	0.786	0.005	0.777	0.005	0.806	0.005
2	1	0.757	0.003	0.748	0.004	0.782	0.003
	2	0.784	0.002	0.777	0.003	0.805	0.002
	3	0.793	0.002	0.785	0.003	0.813	0.001
3	1	0.794	0.002	0.786	0.003	0.814	0.003
	2	0.814	0.003	0.807	0.003	0.832	0.003
	3	0.819	0.003	0.811	0.003	0.837	0.004
4	1	0.864	0.001	0.859	0.001	0.878	0.001
	2	0.872	0.001	0.867	0.000	0.884	0.001
	3	0.873	0.003	0.868	0.003	0.885	0.002

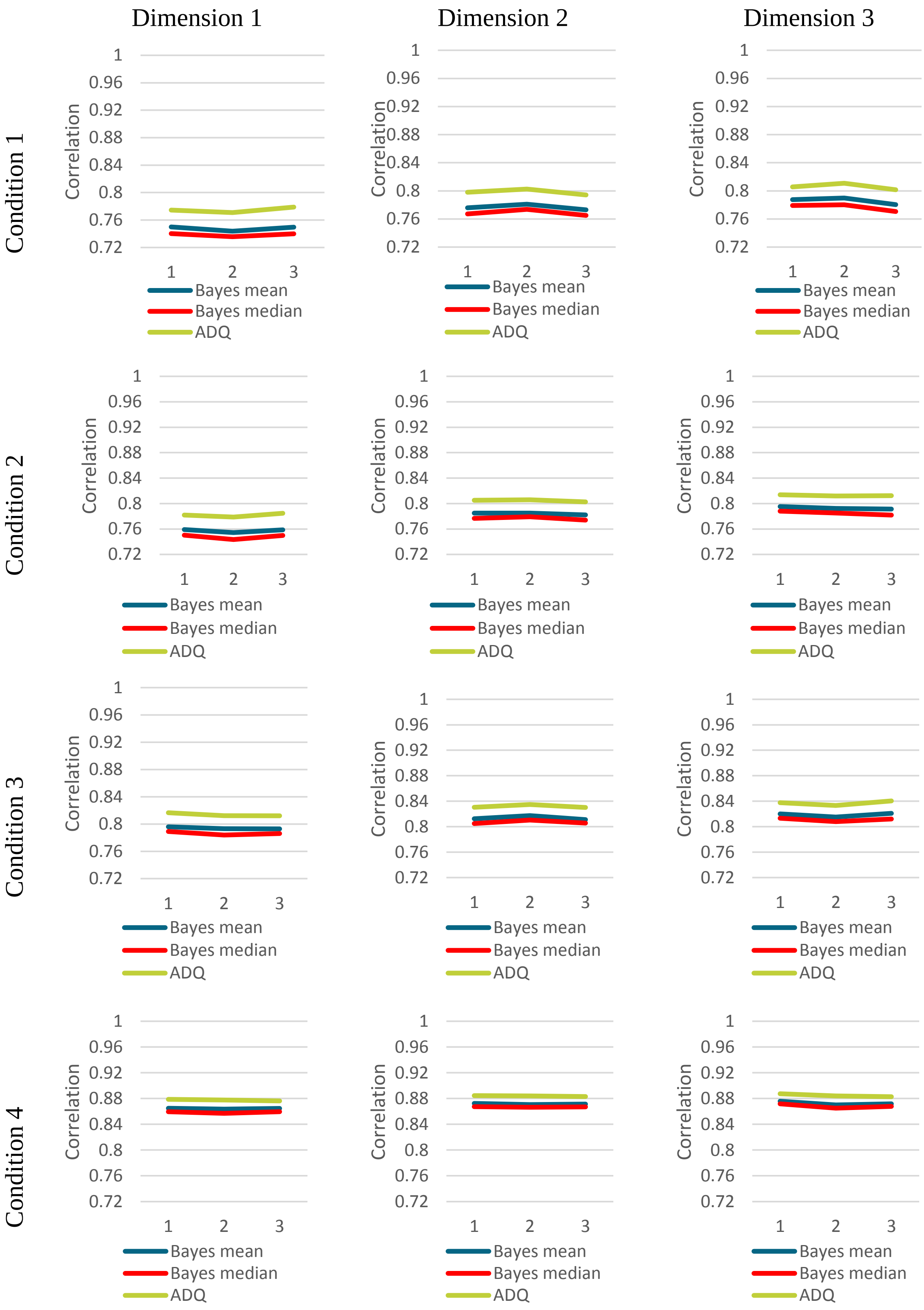

Figure 17. Correlations between true ability and recovered ability for each dimension

Conditions 5 through 8. Conditions 5 through 8 were run using Model 2 and again the only differences were in increase in the correlations between the models. In Model 2, nine items loaded onto dimensions one, two and three, for a total of 27 items that loaded onto only one dimension. Additionally, nine items loaded onto dimensions one and two, one and three and two and three for a total of 27 items loading onto two dimensions. No items loaded onto all three dimensions. In all, 27 items loaded onto each factor. As seen in Model 1, the ADQ estimation technique consistently produced the highest correlations between true ability and the recovered ability and the Bayes median produced the smallest correlations. Likewise, the differences in the recovery correlations were rather small.

As seen in the first model, as the correlation between dimensions increased, the correlation between the recovered abilities and the true abilities values also increased. In addition, Dimension 3 produced the highest recovery correlations for all four conditions under Model 2. Furthermore, the recovery correlations were larger than the recovery correlations seen in the first four conditions even when the correlation between the dimensions were the same. The graphical and tabular results for the first four models are shown in Figure 18 and Table 13.

Table 13.

Average correlation and standard deviations for each dimension for conditions 5-8

Condition	Dimension	Bayes mean r	Bayes mean SD	Bayes median r	Bayes median SD	ADQ r	ADQ SD
5	1	0.784	0.002	0.777	0.003	0.805	0.003
	2	0.805	0.003	0.797	0.004	0.824	0.003
	3	0.838	0.003	0.832	0.003	0.853	0.002
6	1	0.805	0.003	0.799	0.003	0.001	0.001
	2	0.821	0.001	0.815	0.001	0.839	0.000
	3	0.846	0.001	0.841	0.002	0.861	0.001
7	1	0.842	0.002	0.836	0.002	0.856	0.002
	2	0.851	0.002	0.845	0.003	0.867	0.002
	3	0.870	0.002	0.865	0.001	0.882	0.001
8	1	0.912	0.002	0.909	0.003	0.920	0.001
	2	0.916	0.001	0.913	0.001	0.923	0.001
	3	0.920	0.001	0.918	0.001	0.928	0.001

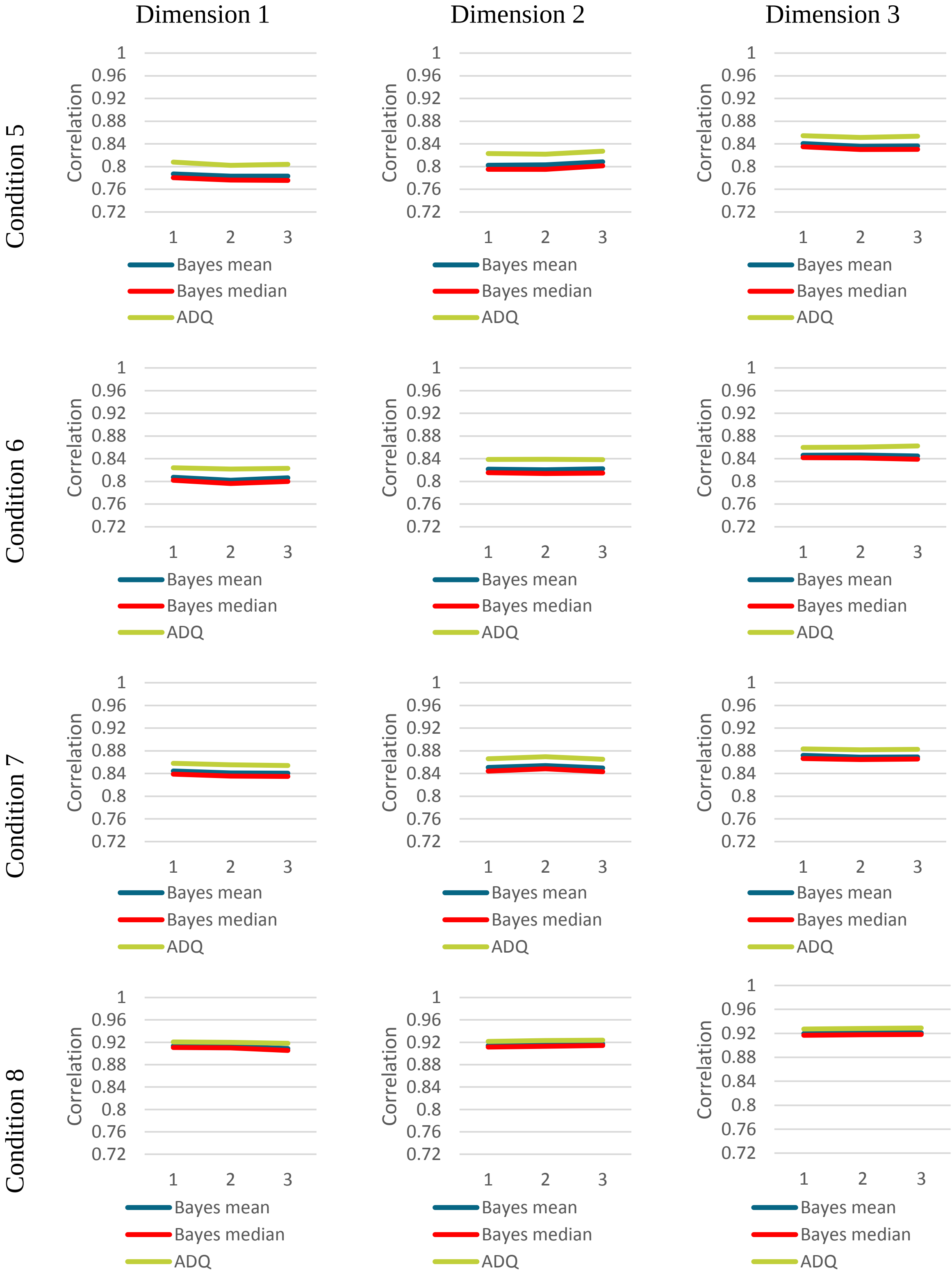

Figure 18. Correlations between true ability and recovered ability for each dimension

Conditions 9 through 12. The third model was used to simulate Conditions 9 through 12 and again the correlations between dimensions increased as the condition number increased. In Model 3, nine items loaded solely onto Dimension 1, nine items solely loaded onto Dimension 2 and nine items solely loaded onto Dimension 3. Of the 54 items, 12 loaded onto two dimension; four loaded onto both dimensions one and two, four loaded onto both dimensions one and three and four loaded onto both dimensions two and three. The remaining 15 items loaded onto all three dimensions. In this model, a total of 32 items loaded onto each dimension, more than in any other model.

As seen in all of the other condition, the ADQ estimation technique performed the best with the highest correlations between recovered ability and the true estimates. The Bayes median had the lowest correlations between the true and the recovered ability levels. Even though the ADQ technique performed better, the differences between the ADQ and the Bayes estimation techniques was the smallest with the third model. For a third time, the recovery correlations for Dimension 3 were higher than that of the other two dimensions. The graphical and tabular results for the first four models are shown in Figure 19 and Table 14.

Table 14.

Average correlation and standard deviations for each dimension for conditions 9-12

Condition	Dimension	Bayes mean r	Bayes mean SD	Bayes median r	Bayes median SD	ADQ r	ADQ SD
9	1	0.750	0.002	0.740	0.000	0.775	0.001
	2	0.778	0.002	0.769	0.003	0.801	0.002
	3	0.827	0.005	0.821	0.005	0.844	0.005
10	1	0.786	0.002	0.779	0.003	0.805	0.001
	2	0.805	0.002	0.798	0.002	0.823	0.001
	3	0.843	0.002	0.837	0.002	0.857	0.003
11	1	0.839	0.002	0.834	0.003	0.853	0.001
	2	0.847	0.000	0.842	0.001	0.862	0.002
	3	0.875	0.003	0.871	0.002	0.887	0.002
12	1	0.919	0.001	0.917	0.001	0.926	0.001
	2	0.921	0.003	0.918	0.003	0.928	0.002
	3	0.928	0.003	0.926	0.003	0.935	0.001

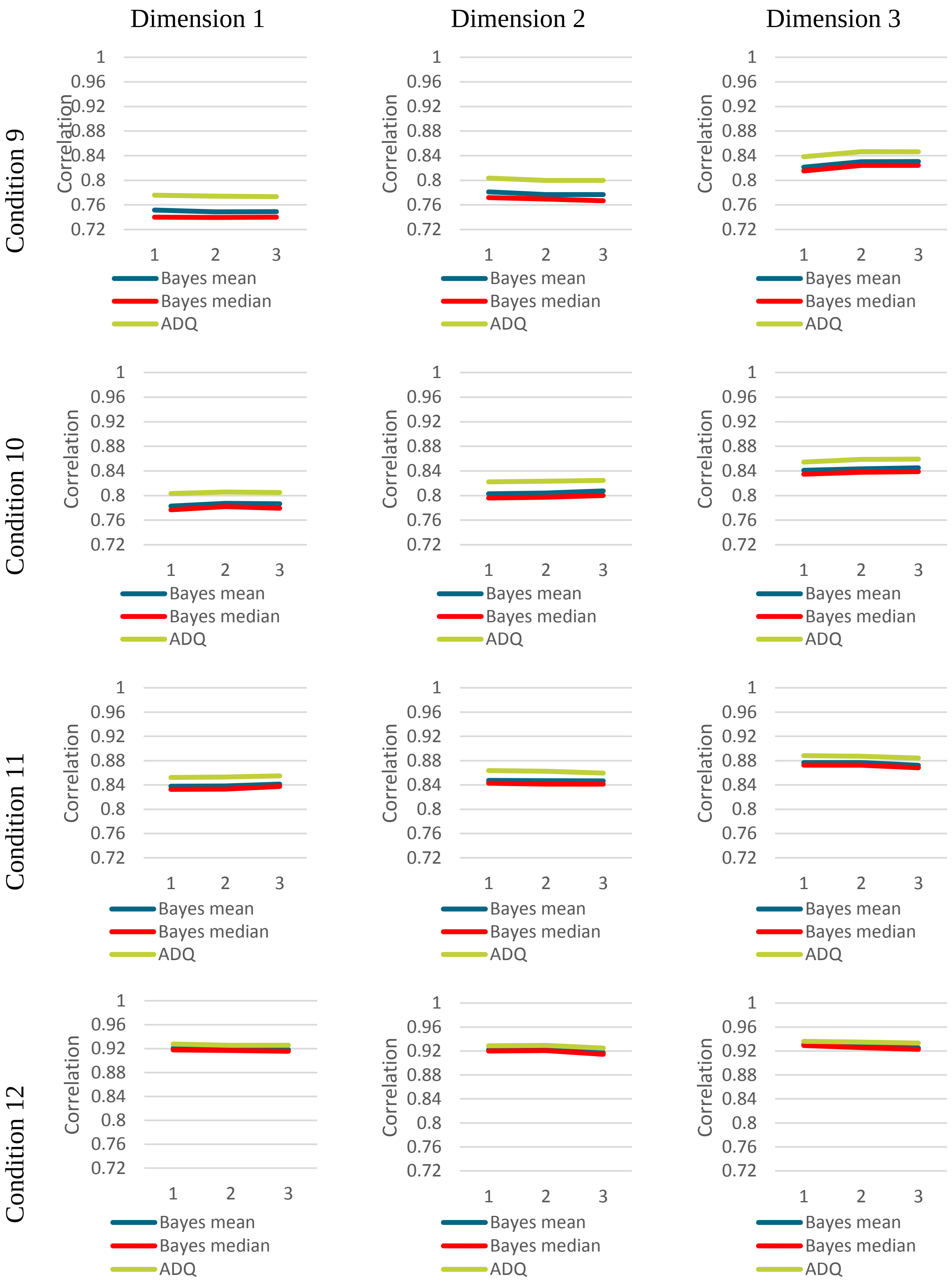

Figure 19. Correlations between true ability and recovered ability for each dimension

5 DISSCUSSION

Each software program provided evidence as to how it would run in the real-world. Even though there was no program that performed the best in all conditions, there was a great deal of information obtained about the efficacy of each estimation technique in a variety of situations.

Item parameters. The overall results of the item parameter recovery are a bit mixed; there is not one program that performed better than the other in all conditions. The Bayes estimation technique consistently produced smaller RMSE values for the MDISC parameter and the MHRM for the *d* parameter. The correlation between the three dimensions and the complexity of the model did not appear to change the results overall. A holistic view shows that the RMSE was generally small for each estimation technique and neither the Bayes or MHRM estimation techniques produced estimates that were far from the true values. It should be noted that these results were the product of only three replications. Even though three replications were used to resemble what practitioners would face in the real-world, it might also mask results that could be seen in the presences of more replications.

IRTPRO 3 and the MHRM estimation technique might prove the slightest bit more beneficial for practitioners because the results are shown using item response theory parameter variables. The Bayes estimation technique in Mplus 7 produced factor loadings and thresholds, which must then be converted to item parameters. Practitioners that are not familiar with factor analysis might not understand how to convert from these parameters to item response parameters, or might do the conversion incorrectly. Even though the Bayes estimation performed better overall, the fact that the item parameters might not be converted correctly and the rather small differences in RMSE suggest that the MHRM estimation technique might be better for practitioners who are interested in recovering MIRT item parameters.

Analogous to the study by Knoll and Berger (1991) this study showed no large differences between a factor analysis approach and an IRT approach to multidimensional test data. Additionally, the results of the current study were similar to the results of Bolt and Lall (2003), Chalmers and Flora (2014), Finch (2010), Han and Paek (2014) and Tate (2003). In those studies, the item parameter recovery appears to be unaffected by the simulation conditions such as change in the correlations between dimensions, sample sizes, and test lengths. Finch (2011) found higher bias in items that loaded onto more than one dimension and when the intercorrelations between dimensions increased. Although the current study did not identify individual item error, the results did parallel Finch and produced larger errors for the MDISC parameter under more complex test models with items loading onto at most two dimensions.

The results of item parameter recovery from Babcock 2011 study and the Chalmers and Flora 2014 showed that highly correlated dimensions produced poorly recovered *a* parameters. The current study showed that the increase in correlations did not have an impact on the recovery of both the MDISC and *d* parameters. Both study results were based on noncompensatory models whereas this study was based on compensatory models. This could account for the difference in the estimation of the *a* parameter and could be a subject for further investigation.

Examinee ability. One noted result was the recovered correlations in Dimension 3. This is an interesting result because the same number of items load onto each dimension. It would not be expected that the correlations for Dimension 3 would be that different from the other correlations between dimensions. At this point, no conclusion has been made as to why the correlations for this dimension were so low. Further investigations need to be done to see if the increased number of dimensions produce smaller correlations overall.

The increase in correlations brings the model closer and closer to the unidimensional model, and therefore, the increase in the recovery correlation from one model to the next is not surprising. The increase in recovery correlations might be attributed to the fact that there is more information available to produce an ability estimate under Model 2 than there was in Model 1. Model 2 uses 27 items per dimension as compared to Model 1 in which there were only 18 items per dimension. On the same comparison, when compared to the other two models, Model 3 had more items per dimension with 32 items loading onto each dimension. Due to the increased data available, the correlations between the true ability levels values and the recovered ability estimates is the greatest for Model 3. Condition 12 had the highest correlation between dimensions with a correlation of .9 and produced the highest of all recovery correlations. This could be due to the fact that not only were there more items per dimension, the correlation was so high that it was similar to a unidimensional test were all items load onto a single factor.

The ADQ estimation technique always produced the highest correlation between the true ability values and the recovered ability estimates. The results also showed that as the number of items increased and the correlation between dimensions increased, the ability of both estimation techniques to recover true values also increased. The final recommendation would be to use the ADQ estimation technique with as many items as possible per dimension. Even though the results showed the highest correlations with the most complex models, the recommendation should not imply that the model must be complex with multiple items loading onto multiple dimension, but it is to suggest that the more information available to estimate the examinees ability the better. The recommendation is also mirrored in the results of the number of items per dimension investigation. As the number of items increased, the capability to recover ability estimates also increased. One caveat, however, is when the assessments are used for high stakes purposes. When the test measures three dimensions, the results tend to suggest that practitioners should use at least 10 items per dimension because anything less than that does not produce accurate results.

The ability estimates recovery for the current study produced comparable results to those seen in previous research. The study design of de la Torre and Patz (2005) was similar to the design of this study. While the current study did not increase the number of abilities and used slightly different values for the correlation between the dimensions, the overall results were the same. Both studies found that as the number of items increased, so too did the correlations between the true and recovered ability estimates. The highest correlations found in de la Torre and Patz (2005) were in the condition with five dimensions and interdimensional correlations of 0.9. This results is synonymous with the results of this study where the highest correlations were also found in the condition with interdimensional correlations of 0.9.

Both de la Torre (2009) and Wang, Chen and Cheng (2004) observed that as the test length increased and the correlation between dimensions increased, the recovery of examinee abilities also increased. Furthermore, Wang, Chen and Cheng (2004) found that four to nine items per dimension produced unreliable results. The same conclusion was obtained in the current study where the increase in interdimensional correlations increased the correlations between true examinee ability and recovered examinee ability. Additionally, the current study also showed that less than 14 items per dimensions produced lower recovery correlations for all estimation techniques.

Overall software comparison. IRTPRO 3 and Mplus 7 have two different user interfaces. IRTPRO 3 utilizes both a point and click interface as well as allowing users to write syntax to run models. With Mplus 7, users can only specify models by writing syntax. The ability to write syntax is a plus for both software systems because it allows users to specify the exact model needed as well as allowing for easy updating or changing of models. The point and click interface makes it less intimidating for a person who might not have a great deal of background in programming or writing syntax. Although the syntax for Mplus 7 is not complicated and is very straight forward, users that are not familiar with writing syntax may not appreciate the power of syntax to make models. Nonetheless, the more user friendly interface of IRTPRO 3 does not make IRTPRO 3 the better program in the ease of use category. It would be expected that the target audience of both software programs would have been exposed to writing syntax in some form and therefore having the point and click interface of IRTPRO 3 is not an inherit benefit. IRTPRO 3 could be beneficial as an introduction to item analysis when trying to explain an overview of how the analysis works to those that are not interested in the background calculations.

Mplus 7 does do a better job in communicating program errors. If the model does not run, Mplus 7 does provide information as to where the error occurred. When errors occur in IRTPRO 3, it displayed as an unknown error. The unknown error received in the analysis could be attributed to the fact that the error did not trigger any known error codes in the program; however, the only errors that were received in IRTPRO 3 were unknown errors so no further comparisons could be made. The lack of error explanation can possibly frustrate practitioners who again might be relying on a single dataset to run analyses and would therefore be left with no suggested direction. Both programs ran rather efficiently. The Bayes estimation took the longest to run out of all five programs. Be that as it may, in the most complex models with the largest number of examinees, the increased time it took to run the model in comparison with the other estimation techniques was negligible.

IRTPRO 3 has the ability to run unidimensional and multidimensional models as well as 1PL, 2PL, 3PL, and partial credit models. Mplus 7 can also run both unidimensional and multidimensional models, however some studies have shown that it has difficulty estimating the pseudo-guessing (*c*) parameter (Finch, 2010; Tate, 2003). Additionally, newer versions of Mplus produce both factor loadings and thresholds as well as item difficulty and discrimination and therefore no converting is necessary. Both software programs can run simulations and can produce graphics for items and tests. Mplus 7 can also be used to run both confirmatory factor analysis and exploratory factor analysis. IRTPRO 3 cannot run these analyses and therefore the user must know a priori which items load onto which dimensions when modeling a multidimensional test. Not being able to run an EFA indicates that IRTPRO 3 must be used in conjunction with another software tool that has the ability to do such.

The ability of IRTPRO 3 to run a variety of models and even different models in different test sections makes it far superior to Mplus 7 in modeling capabilities. The ability of Mplus 7 to run both exploratory factor analysis and confirmatory factor analysis provides benefits to practitioners who want to use or can only afford to purchase one software application. Given such, the intended use of the software and the needs of the testing program have to be factored in when deciding which is the best software. Figure 20 provides a visual comparison of each software program as well as other information.

Table 15.

IRT Software comparison

	IRTPRO 3	Mplus 7
Ability to write syntax	yes	yes
Point and click user interface	yes	no
1PL models	yes	yes
2PL models	yes	yes
3PL models	yes	yes
Exploratory factor analysis	no	yes
Confirmatory factor analysis	no	yes
Provides item parameters	yes	Yes, but must convert
Provides examinee ability	yes	yes
Graphics	yes	yes

Overall estimation comparison. The WLSMV estimation technique was shown to have issues with convergence under the default settings. When the WLSMV converged, it produced moderately stable estimates in most conditions. One noticeable difference was when the WLSMV estimator was used to recover true examinee abilities. In this instance, not only did the WLSMV technique have an issue with convergence, it also produced a negative correlation in recovering true examinee abilities. As such, the WLSMV technique was not chosen in any part of this study for further investigation. The Mplus syntax can be amended to help with convergence issues when using the WLSMV estimator and research has shown that when the model converges, the WLSMV estimator preforms very well (Beauducel & Herzberg, 2006; Muthén, Du Toit, & Spisic, 1997; Yu, 2002). Due to the results seen in this investigation, it would be advisable to choose another estimation technique over the WLSMV estimation technique in the practical setting if the practitioner is not well versed in Mplus syntax and is not confident in their ability to modify syntax when models do not converge.

Depending on the test in question, a practitioner might have limited data access needed to calculate examinee ability and item parameters. In the case of WLSMV, this is an issue when the file that the practitioner has does not converge when using that estimation technique. The WLSMV estimations techniques inability to converge was even seen at the larger sample sizes. If the WLSMV estimation technique does not converge, then the practitioner will have absolutely no data to use and thus make any item analyses irrelevant.

The Bayes estimation technique, on the other hand, did a suitable job in recovering both item parameter estimates as well as ability estimates. Even though it did not perform as well as the other estimation techniques studied, the results were relatively close to the true values. When using Bayesian estimation in Mplus 7, the practitioner has the ability to specify the required number of imputations. The more imputations requested, the longer it takes for the program to run. The increased run time is a possible limitation of this technique, but, the limitation can be mitigated by changing the number of imputations. Additionally, the Bayes estimation technique produced fairly sensible estimates when sample sizes were on the smaller side. For these reasons, practitioners who choose Mplus 7 might want to use the Bayes estimation technique for all item response theory analyses.

For IRTRPO 3, the BA estimation technique appears to be the inferior of the three techniques studied. BA was selected as the inferior due to the lack of estimation results. When the BA estimation technique was used with models that had more than one dimension loading onto an item, no results were given. The software would only produce a message that stated there was an unknown error. Nonetheless, when a simple between-item structure was used, the BA estimation technique performed admirably. Practitioners using this estimation technique would be limited to simple test structures. IRTPRO 3, however, does provide the ability to use other estimation techniques in the presence of more complex models.

The ADQ estimation technique did not produce any errors in any of the simulation conditions. The recovery results varied the most out of all estimation techniques. A good example of this was in the identification of an adequate sample size. Although the ADQ estimation technique converged, the RMSE varied significantly depending on the number of examinees. It would be understandable for the RMSE to decrease greatly as the number of examines increased; however, this was not the case. The RMSE both increased and decreased as the number of examinees increased. The ADQ estimation technique did show to be the best technique when it came to recovering examinee abilities. The correlations for ADQ were consistently higher than that of the Bayes estimation technique.

As stated previously, the literature suggested the ADQ estimation technique works better in more complex models and models with higher dimensions (Cai, Thissen, & du Toit, 2015a; Schilling & Bock, 2005). A few of the conditions in the item parameter recovery section of the study were rerun using the ADQ estimation technique. The RMSE for the ADQ estimation technique was not noticeably smaller than that of the other techniques explored.

Finally, the MHRM estimation technique also suffered from issues that other estimations techniques did not encounter. One noticeable issue was that of ability recovery. Of all estimation techniques, MHRM unfailingly produced the lowest overall correlations for each condition and in each dimension. Not only were the correlations the lowest out of all five estimation techniques, the correlations were consistently in the low to moderate level between 0.35 and 0.48 while all other estimation techniques were in the 0.60 range.

Given the benefits and issues seen in the IRTPRO 3 estimation techniques, it is hard to identify one overall estimation technique that should be used in all situations. If a practitioner is using a simple structure and only wants to use one estimation technique for both item parameter and examinee ability analysis, the BA estimation technique would be the best. If a practitioner has any complexity in the model and is only interested in the item parameters, it would be beneficial to use the MHRM technique. Finally, if the practitioner is interested in examinee ability alone, the ADQ technique would be the best.

Overall, the capability of IRTPRO 3 to use different analyses depending on the desired information, the capacity to produce item parameters without the need to convert from lambdas and thresholds to discrimination and difficulties, and the ability to model a variety of IRT test structures ultimately gives IRTPRO 3 a slight advantage over Mplus 7.

Limitations. Due to the limited resources available for this study, there are a few limitations that should be acknowledge. Although they might not have a profound impact on the overall results, they could potentially impact future research. The limitations include the number of repetitions, the models used, and how the estimation techniques were run.

As stated in numerous instances, the intended audience for this inquiry was practitioners that have might have access to limited amounts of data. The current study was limited to only three replications in order to imitate what may perhaps be encountered in the real world. Even though this was done with the practitioner in mind, it is also a limitation. The limited number of data sets might have masked an issue that could be seen with more data or it could have pinpointed a difference that seemed rather large though would appear smaller in the presence of multiple samples.

Another possible limitation is the models that were used in the analysis. Two complex and one simple three dimensional models were used in this study. While these models were of interest to the researcher, they are, by far, not the only models available. The inclusion of different types of models as well as the inclusion of a variety of dimensions could help illustrate other issues or benefits related to the recovery of item parameters and ability estimates.

Each estimation technique could have been run in a variety of ways. For instance, the number of quadrature points were selected based on research, but, the number of quadrature points could have been changed. The research indicates that two to three quadrature points should be used per dimension, and with a three-dimensional test, the number of quadrature points recommended could range from six to nine. The selection of different quadrature points for the BA and ADQ estimation technique could have changed the results of the study.

Future research. No study could possibly address all questions related to a single topic. Given the limitations elucidated in the previous section, the direction of the future research should be such that addresses the limitations. Future research could provide practitioners with a more robust assessment of the estimation techniques investigated in this study as well as other estimation techniques available for item parameter analyses. The future directions should relate to the number of repetitions, estimation models, other software programs, and the use of real data.

As stated in the limitation section, one important limitation is the use of only three replications. Future research should involve more replications in order to better demonstrate the abilities of each estimation technique. The increase in replications could be very beneficial in identifying an adequate sample size. The analysis conducted in the current study showed that the ADQ estimation technique was very unstable; though, more replications could smooth out the results of the estimation technique and could be used to make more concrete suggestions as to the number of examinees needed. Additionally, the increase in repetitions could be helpful in understanding how the number of items per dimension impacts the ability to recover examinee estimates. Again, real world scenarios were the bases of the current analysis. However, a more in depth simulation study could help inform practice in a different way.

Future research should also center around the models used and the software program employed. At the conclusion of this study, Scientific Software International released IRTPRO 4. The new software updates might change some of the results seen in the current investigation. Additionally, other software such as SAS/STATA®, R (2013) and Stata 14 (StataCorp 2015), provide the opportunity to use different models including 1PL, 3PL, graded response, and compensatory models as well as other test designs.

Moreover, the inclusion of real-world data could shape the direction of future research. All results of the current exploration focused solely on simulated data. The inclusion of real world data could provide practitioners as well as researchers with valuable information. Studies could incorporate the use of EFA to identify the items and which dimensions they map onto, as well as a follow up item analyses to identify item parameters and ability estimations.

Finally, this research could be expanded to the realm of practice based professional development. Practice based professional development is focused on the development of the practitioner’s understandings of the skills needed for an effective education practice as opposed to the knowledge of the practice (Harris et al., 2012). The current study could help shape district and large school assessment designers understanding of the practical difference between IRT and factor analysis estimation procedures for multidimensional test. The information presented in this study as well as other information related to test design can be developed into professional learning activities that could help practitioners incorporate better assessments at all levels of education.

Though there are limitations in the study and future directions could extend the knowledge available on the given topic, the current study contributes to the realm of item response theory. The current study provides educational professions with necessary information to make informed decisions as to which analyses and estimation techniques to use under a variety of circumstances.

REFERENCES

Ackerman, Terry A. "Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items." *Applied Psychological Measurement* 13.2 (1989): 113-127.

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

Babcock, B. (2011). Estimating a Noncompensatory IRT Model Using Metropolis Within Gibbs Sampling. *Applied Psychological Measurement*, *35*(4), 317-329. doi:10.1177/0146621610392366

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika*, *48*(4), 493-517.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bentler, P.M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588.

Bo, Z., & Stone, C. A. (2008). Evaluating Item Fit for Multidimensional Item Response Models. *Educational & Psychological Measurement*, *68*(2), 181-196.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.

Bock, R. D. (1985). Contributions of empirical Bayes and marginal maximum likelihood methods to the measurement of individual differences. *Measurement and personality assessment*, 75-99.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303-316.

Bolt, D. M. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395-414.

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.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307-335.

Cai, L., Thissen, D., & du Toit, S. H. C. (2015a). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Cai, L., Thissen, D., & du Toit, S. H. C. (2015b). IRTPRO Users Guide: Scientific Software International.

Cai, L., & Wirth, R. J. (2013). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring.

Carroll, J. D., Williams, B. A., & Levine, M. V. (2007). Multidimensional modeling with unidimensional approximations. *Journal of Mathematical Psychology, 51*(4), 207-228.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

De Ayala, R. J. (2013). *Theory and practice of item response theory*. Guilford Publications.

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, *33*(6), 465-485.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295-311.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495-515.

Finch, H. (2010). Item Parameter Estimation for the MIRT Model Bias and Precision of Confirmatory Factor Analysis—Based Models. *Applied Psychological Measurement*, *34*(1), 10-26.

Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*,*71*(4), 663-683.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, *42*(2), 149-169.

Fraser, C., & McDonald, R. P. (1988). NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. *The University of New England, Armidale, Australia*.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Measurement methods for the social sciences series, Vol. 2).

Han, K. T., & Paek, I. (2014). A Review of Commercial Software Packages for Multidimensional IRT Modeling. *Applied Psychological Measurement*, *38*(6), 486-498.

Harris, K. R., Lane, K. L., Graham, S., Driscoll, S. A., Sandmel, K., Brindle, M., & Schatschneider, C. (2012). Practice-based professional development for self-regulated strategies development in writing: A randomized controlled study. *Journal of Teacher Education*, *63*(2), 103-119.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). Applied statistics for the behavioral sciences (5th ed).

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Articles*, 2.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Iglewicz, B., Hoaglin, D. How to detect and handle outliers. ASQC Quality Press, 1993.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data* (Vol. 1). SAGE Publications, Incorporated.

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1984). LISREL VI user's guide (3rd ed.). Mooresville, IN: Scientific Software.

Kahraman, N. (2013). Unidimensional Interpretations for Multidimensional Test Items. Journal Of Educational Measurement, 50(2), 227-246.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136-153.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*(4), 587-599.

Kim, S. Y., Suh, Y., Kim, J. S., Albanese, M. A., & Langer, M. M. (2013). Single and Multiple Ability Estimation in the SEM Framework: A Noninformative Bayesian Estimation Approach. *Multivariate behavioral research*, *48*(4), 563-591.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate behavioral research*, 26(3), 457-477.

Linacre, J. M. (2010). Winstep-Rasch model computer program. Version 3.69. 1.16.

Lord, F. M., Novick, M. R., (1968). Statistical theories of mental test scores.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372-379.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49-57.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods*, *15*(3), 389-390.

Mulaik, S. A. (1972). A mathematical investigation of some multidimensional Rasch models for psychological tests. In annual meeting of the Psychometric Society, Princeton, NJ.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551-560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*,*49*(1), 115-132.

Muthén, B., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-timepoint example. *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*, 1-29.

Muthén, B., Du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*, *75*(1), 1-45.

Muthen, B. O., Muthen, L. K., & Asparouhov, T. (2015). Estimator choices with categorical outcomes. Retrieved from https://www.statmodel.com/download/EstimatorChoices.pdf

Muthén, L. K., & Muthén, B. O. (2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289-298.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, 1(2), 66-82.

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

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

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement,15*(4), 361-373.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533-555.

Segall, D. O. (1998). IFACT computer program Version 2.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [Computer program]. *Seaside, CA: Defense Manpower Data Center*.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*(1), 79-97.

StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

Statistical Analysis System Institute. (1999). *SAS/STAT Computer software, version 8* (Vol. 2). SAS Institute.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

Sympson, J. B. (1978). A model for testing with multidimensional items. In*Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structural models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, 42, 233-239.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159-203.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10.

Wang, C., & Nydick, S. W. (2015). Comparing Two Algorithms for Calibrating the Restricted Non-Compensatory Multidimensional IRT Model. *Applied Psychological Measurement*, *39*(2), 119-134.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295-316

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological methods*, *9*(1), 116.

Wen-Chung, W. (2004). Direct Estimation of Correlations as a Measure of Association Strength using Multidimensional Item Response Models. *Educational & Psychological Measurement*, *64*(6), 937-955.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests .*Psychometrika*, *45*(4), 479-494.

Wilson, D. T., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. SSI, Scientific Software International.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ACER ConQuest 2.0: General item response modelling software [computer program manual]. *Camberwell, Australia: Australian Council for Educational Research*.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245-262.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79-94.

Yu, C.Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Unpublished doctoral dissertation, University of California, Los Angeles.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). BILOG-MG3 [Computer software]. Chicago: Scientific Software International.

APPENDICES

Appendix A Item Parameters for item parameter recovery

Item parameters for the between and within-item multidimensional IRT model
Model 2 item parameter recovery

Item number	a_1	a_2	a_3	d
1	0.04	0.00	0.00	0.08
2	0.02	0.00	0.00	-0.09
3	0.15	0.00	0.00	-0.29
4	0.09	0.00	0.00	-0.29
5	0.16	0.00	0.33	-0.22
6	0.16	0.00	0.67	-0.31
7	0.20	0.00	0.89	-0.63
8	0.12	0.00	0.51	0.40
9	0.12	0.04	0.00	0.04
10	0.06	0.05	0.00	0.05
11	0.47	0.14	0.00	0.45
12	0.63	0.02	0.00	-0.14
13	0.00	0.42	0.00	0.23
14	0.00	0.19	0.00	-1.02
15	0.00	0.51	0.00	0.23
16	0.00	0.55	0.00	-0.16
17	0.00	0.05	0.05	-0.16
18	0.00	0.16	0.12	0.31
19	0.00	0.18	0.06	0.71
20	0.00	0.22	0.05	0.03
21	0.00	0.00	0.05	0.38
22	0.00	0.00	0.11	-0.12
23	0.00	0.00	0.16	-0.17
24	0.00	0.00	0.12	-1.07

Item parameters for the complex between and within-item multidimensional IRT model

Model 3 item parameter recovery

Item number	a_1	a_2	a_3	d
1	0.04	0.00	0.00	0.08
2	0.02	0.00	0.00	-0.09
3	0.15	0.00	0.00	-0.29
4	0.19	0.03	0.00	-0.29
5	0.16	0.10	0.00	-0.22
6	0.16	0.16	0.67	-0.31
7	0.20	0.03	0.00	-0.63
8	0.12	0.12	0.51	0.40
9	0.00	0.06	0.00	0.04
10	0.00	0.47	0.00	0.05
11	0.00	0.63	0.00	0.45
12	0.00	0.04	0.00	-0.14
13	0.00	0.03	0.11	0.23
14	1.09	0.02	0.17	-1.02
15	0.00	0.01	0.08	0.23
16	0.55	0.00	0.05	-0.16
17	0.00	0.00	0.85	-0.16
18	0.00	0.00	0.54	0.31
19	0.00	0.00	0.56	0.71
20	0.00	0.00	0.39	0.03
21	0.05	0.00	0.06	0.38
22	0.11	0.27	0.11	-0.12
23	0.16	0.00	0.15	-0.17
24	0.12	0.18	0.01	-1.07

Appendix B Item Parameters for item parameter recovery

Item parameters for the between-item multidimensional IRT model

Model 1 ability estimation recovery

Item Number	*a*1	*a*2	*a*3	*d*	Item Number	*a*1	*a*2	*a*3	*d*
1	0.50	0.00	0.00	0.08	28	0.00	0.46	0.00	-0.49
2	0.27	0.00	0.00	-0.09	29	0.00	0.38	0.00	-0.36
3	0.58	0.00	0.00	-0.29	30	0.00	1.23	0.00	-0.73
4	0.67	0.00	0.00	-0.29	31	0.00	0.98	0.00	-0.83
5	0.33	0.00	0.00	-0.22	32	0.00	0.66	0.00	0.19
6	0.67	0.00	0.00	-0.31	33	0.00	0.76	0.00	-0.22
7	0.89	0.00	0.00	-0.63	34	0.00	0.71	0.00	-0.20
8	0.51	0.00	0.00	0.40	35	0.00	0.68	0.00	0.43
9	0.60	0.00	0.00	0.04	36	0.00	0.93	0.00	-0.16
10	0.54	0.00	0.00	0.05	37	0.00	0.00	0.59	0.55
11	0.47	0.00	0.00	0.45	38	0.00	0.00	0.53	-0.03
12	0.63	0.00	0.00	-0.14	39	0.00	0.00	0.54	-0.17
13	0.42	0.00	0.00	0.23	40	0.00	0.00	0.29	-0.39
14	1.09	0.00	0.00	-1.02	41	0.00	0.00	0.61	-0.78
15	0.51	0.00	0.00	0.23	42	0.00	0.00	0.81	-1.67
16	0.55	0.00	0.00	-0.16	43	0.00	0.00	0.39	-0.47
17	0.85	0.00	0.00	-0.16	44	0.00	0.00	0.67	-0.26
18	0.54	0.00	0.00	0.31	45	0.00	0.00	0.68	-0.86
19	0.00	0.37	0.00	-0.14	46	0.00	0.00	0.29	-0.10
20	0.00	0.73	0.00	0.29	47	0.00	0.00	0.32	-0.24
21	0.00	0.91	0.00	-0.21	48	0.00	0.00	0.54	-0.52
22	0.00	0.35	0.00	0.16	49	0.00	0.00	0.79	0.78
23	0.00	0.30	0.00	0.01	50	0.00	0.00	1.11	-0.19
24	0.00	0.65	0.00	0.06	51	0.00	0.00	0.69	0.44
25	0.00	0.52	0.00	-0.08	52	0.00	0.00	1.73	-1.62
26	0.00	0.31	0.00	-0.26	53	0.00	0.00	0.80	0.40

27	0.00	0.64	0.00	-0.43	54	0.00	0.00	0.91	-0.32

Item Parameters for the between and within-item multidimensional IRT model
Model 2 ability estimation recovery

Item Number	*a*1	*a*2	*a*3	*d*	Item Number	*a*1	*a*2	*a*3	*d*
1	0.50	0.00	0.00	0.08	28	0.54	0.46	0.00	-0.49
2	0.27	0.00	0.00	-0.09	29	0.47	0.38	0.00	-0.36
3	0.58	0.00	0.00	-0.29	30	0.63	1.23	0.00	-0.73
4	0.67	0.00	0.00	-0.29	31	0.42	0.98	0.00	-0.83
5	0.33	0.00	0.00	-0.22	32	1.09	0.66	0.00	0.19
6	0.67	0.00	0.00	-0.31	33	0.51	0.76	0.00	-0.22
7	0.89	0.00	0.00	-0.63	34	0.55	0.71	0.00	-0.20
8	0.51	0.00	0.00	0.40	35	0.85	0.68	0.00	0.43
9	0.60	0.00	0.00	0.04	36	0.54	0.93	0.00	-0.16
10	0.00	0.37	0.00	0.05	37	0.42	0.00	0.93	0.55
11	0.00	0.73	0.00	0.45	38	1.09	0.00	0.58	-0.03
12	0.00	0.91	0.00	-0.14	39	0.51	0.00	1.53	-0.17
13	0.00	0.35	0.00	0.23	40	0.55	0.00	0.69	-0.39
14	0.00	0.30	0.00	-1.02	41	0.85	0.00	0.79	-0.78
15	0.00	0.65	0.00	0.23	42	0.54	0.00	1.06	-1.67
16	0.00	0.52	0.00	-0.16	43	0.56	0.00	0.75	-0.47
17	0.00	0.31	0.00	-0.16	44	0.39	0.00	0.32	-0.26
18	0.00	0.64	0.00	0.31	45	0.65	0.00	0.54	-0.86
19	0.00	0.00	0.59	-0.14	46	0.00	0.65	0.29	-0.10
20	0.00	0.00	0.53	0.29	47	0.00	0.27	0.32	-0.24
21	0.00	0.00	0.54	-0.21	48	0.00	0.65	0.54	-0.52
22	0.00	0.00	0.29	0.16	49	0.00	0.18	0.79	0.78
23	0.00	0.00	0.61	0.01	50	0.00	0.98	1.11	-0.19
24	0.00	0.00	0.81	0.06	51	0.00	0.66	0.69	0.44
25	0.00	0.00	0.39	-0.08	52	0.00	0.76	1.73	-1.62
26	0.00	0.00	0.67	-0.26	53	0.00	0.71	0.80	0.40
27	0.00	0.00	0.68	-0.43	54	0.00	0.68	0.91	-0.32

Item parameters for the complex between and within-item multidimensional IRT model Model 3 ability estimation recovery

Item	*a*1	*a*2	*a*3	*d*	Item	*a*1	*a*2	*a*3	*d*
Number					Number				
1	0.50	0.00	0.00	0.08	28	0.54	0.46	0.00	-0.49
2	0.27	0.00	0.00	-0.09	29	0.47	0.38	0.00	-0.36
3	0.58	0.00	0.00	-0.29	30	0.63	1.23	0.00	-0.73
4	0.67	0.00	0.00	-0.29	31	0.42	0.98	0.00	-0.83
5	0.33	0.00	0.00	-0.22	32	1.09	0.66	0.71	0.19
6	0.67	0.00	0.00	-0.31	33	0.51	0.76	1.23	-0.22
7	0.89	0.00	0.00	-0.63	34	0.55	0.71	0.98	-0.20
8	0.51	0.00	0.00	0.40	35	0.85	0.68	0.66	0.43
9	0.60	0.00	0.00	0.04	36	0.54	0.93	0.76	-0.16
10	0.00	0.37	0.00	0.05	37	0.42	0.00	0.93	0.55
11	0.00	0.73	0.00	0.45	38	1.09	0.00	0.58	-0.03
12	0.00	0.91	0.00	-0.14	39	0.51	0.00	1.53	-0.17
13	0.00	0.35	0.00	0.23	40	0.55	0.00	0.69	-0.39
14	0.00	0.30	0.00	-1.02	41	0.85	0.65	0.79	-0.78
15	0.00	0.65	0.00	0.23	42	0.54	0.52	1.06	-1.67
16	0.00	0.52	0.00	-0.16	43	0.56	0.31	0.75	-0.47
17	0.00	0.31	0.00	-0.16	44	0.39	0.64	0.32	-0.26
18	0.00	0.64	0.00	0.31	45	0.65	0.46	0.54	-0.86
19	0.00	0.00	0.59	-0.14	46	0.00	0.65	0.29	-0.10
20	0.00	0.00	0.53	0.29	47	0.00	0.27	0.32	-0.24
21	0.00	0.00	0.54	-0.21	48	0.00	0.65	0.54	-0.52
22	0.00	0.00	0.29	0.16	49	0.00	0.18	0.79	0.78
23	0.00	0.00	0.61	0.01	50	0.37	0.98	1.11	-0.19
24	0.00	0.00	0.81	0.06	51	0.73	0.66	0.69	0.44
25	0.00	0.00	0.39	-0.08	52	0.91	0.76	1.73	-1.62
26	0.00	0.00	0.67	-0.26	53	0.35	0.71	0.80	0.40
27	0.00	0.00	0.68	-0.43	54	0.30	0.68	0.91	-0.32