ACCURACY OF GLOBAL FIT INDICES AS INDICTORS OF MULTIDIMENSIONALITY IN MULTIDIMENSIONAL RASCH ANALYSIS

Leigh M. Harrell

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Educational Research and Evaluation

Edward W. Wolfe
Gary Skaggs
Yasuo Miyazaki
George Terrell

October 28, 2009
Blacksburg, VA

Keywords: Rasch analysis, confirmatory factor analysis, AIC, BIC, global fit indices, model selection

ACCURACY OF GLOBAL FIT INDICES AS INDICTORS OF MULTIDIMENSIONALITY IN MULTIDIMENSIONAL RASCH ANALYSIS

Leigh Harrell

ABSTRACT

Most research on confirmatory factor analysis using global fit indices (AIC, BIC, AICc, and CAIC) has been in the structural equation modeling framework. Little research has been done concerning application of these indices to item response models, especially within the framework of multidimensional Rasch analysis. The results of two simulations studies that investigated how sample size, between-dimension correlation, and test length affect the accuracy of these indices in model recovery using a multidimensional Rasch analysis are described in this dissertation. The first study analyzed dichotomous data, with model-to-data misfit as an additional independent variable. The second study analyzed polytomous data, with rating scale structure as an additional independent variable. The interaction effect between global fit index and between-dimension correlation had very large effect sizes in both studies. At higher values of between-dimension correlation, AIC indicated the correct two-dimension generating structure slightly more often than does the BIC or CAIC. The correlation by test length interaction had an odds ratio indicating practical importance in the polytomous study but not the dichotomous study. The combination of shorter tests and higher correlations resulted in a difficult-to-detect distinction being modeled with less statistical information. The correlation by index interaction in the dichotomous study had an odds ratio indicating practical importance. As expected, the results

demonstrated that violations of the Rasch model assumptions are magnified at higher between-dimension correlations. Recommendations for practitioners working with highly correlated multidimensional data include creating moderate length (roughly 40 items) instruments, minimizing data-to-model misfit in the choice of model used for confirmatory factor analysis (MRCMLM or other MIRT models), and making decisions based on multiple global indices instead of depending on one index in particular.

Dedication

I dedicate this work to my parents and other family members, who taught me that I should set goals and work hard to reach them. I am grateful to them for their support and for instilling in me a love of learning that kept me going through the rougher moments of graduate school.

Acknowledgements

I would like to thank my family and friends, the faculty and my classmates in the Educational Research Evaluation program, my colleagues in the Statistics department, and my fellow CAUSEmos Statistics Education researchers. Without their encouragement and understanding, I would not have been able to earn my PhD while working full-time as an instructor and an advisor.

I would like to acknowledge the guidance I received from my mentors throughout my journey through education: Mrs. Donna Davis at St. Clare; Dr. Bob Albertson, Dr. Margaret Zimmerman and Mrs. Deborah York at Virginia Wesleyan; Dr. Lynne Seymour and Dr. Mary Meyers in the Department of Statistics at University of Georgia; Dr. Edward Wolfe and Dr. Elizabeth Creamer in the Educational Research and Evaluation program at Virginia Tech. I would also like to acknowledge the other member of my doctoral committee: Dr. Gary Skaggs, Dr. Yasuo Miyazaki, and Dr. George Terrell. To all of you, your guidance, confidence in me, and patience with me made the difference in completing this journey.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

Confirmatory factor analysis (CFA) is frequently used to evaluate the theoretical dimensions of a measurement instrument. Both the number of factors and the pattern of how the items correspond to the factors (factor loadings) are verified through CFA. The results of such analyses serve as structural validity evidence for scores based on responses to the instrument.

Currently, CFA is conducted almost exclusively within an SEM framework. In this framework, when the observed indicators of the factors are continuous, the variance-covariance matrix of the observed data (S) is compared to the estimated variance-covariance matrix (Σ) from the specified factor structure to determine how well the proposed model fits the data. When the observed indicators are not continuous, the estimation procedure is handled slightly differently and the input matrix, S, becomes a correlation matrix. In this SEM framework, referred to as the threshold model, y* reflects the amount of an underlying continuous and normally distributed characteristic that is required to respond in a certain category of the observed categorical variable (Takane & De Leeuw, 1987; Knol & Berger, 1991). Binary outcomes have one threshold (τ_i) that is the point on y* where y=1 if the threshold is exceeded. Polytomous outcomes will have one less threshold than the number of categories (Brown, 2006).

With the development of multidimensional item response (MIRT) models, it is possible for confirmatory factor analysis to be conducted within a latent trait test theory framework. In general, latent trait models simultaneously estimate where an examinee is located along the continuum of the trait being measured (θ_n) and item difficulty (b_i). Responses are assumed to be either dichotomously or polytomously scored. In a Rasch framework, the slopes of the item

characteristic curves (α_i) are assumed to be equal to one, whereas two-parameter and three-parameter models allow for the items to have different slopes. It has been shown that for dichotomous items, the difficulty parameters (b_i) and the slope parameters (α_i) from a multidimensional two-parameter normal ogive model can be expressed as functions of the parameters of the factor analysis threshold model: $b_i = \lambda_i/\psi_I$ and $\alpha_I = \tau_i/\psi_i$ (Takane & De Leeuw, 1987; Knol & Berger, 1991).

In CFA, it is common to compare the fit of multiple latent trait models, using one of two methods to compare the relative fit of those models: hypothesis tests, such as a chi-squared test, and global fit indices, including the Akaike Information Criteria (AIC; Akaike, 1973). The chi-squared test is known to be sensitive to sample size and results in a statistical significance even when the differences between the estimated and observed covariance matrices are negligible (Brown, 2006). Hence, the use of fit indices is prevalent in CFA.

There are numerous fit indices that are used in CFA in SEM applications. The performance of these indices has been the focus of several studies (Cudeck & Browne, 1983; Haughton et al., 1997; Marsh et al., 1988; Bandalos, 1993; Hu & Bentler, 1998; Whittaker & Stapleton, 2006). Conditions such as sample size, model misspecification, and estimation method have been investigated. Many studies have shown that sample is influential in the performance of fit indices.

Some of the fit indices used in SEM, such as AIC and BIC, are utilized in confirmatory factor analysis using multidimensional item response models. However, there is little research on the performance of these fit indices in CFA using this framework. This study seeks to examine the accuracy of the global fit indices (AIC, AICc, CAIC, and BIC) in a multidimensional Rasch analysis for dichotomously-scored data generated from the multidimensional item response

models, under conditions of short and moderate length tests, small to moderate sample sizes, and moderate to high correlation between dimensions on a simulated two-dimensional instrument. The study also examines the effects of model-to-data misfit on the accuracy of global fit indices as applied to CFA within an MIRT framework. Additionally, the accuracy of the global fit indices in an analysis using polytomous data will be explored, while varying test length, sample size, between-dimension correlation, and rating scale structure.

RESEARCH QUESTIONS

The study seeks to address the following questions:

(1) How do test length, sample size, and between-dimension correlation affect the accuracy with which four global fit indices (AIC, AICc, CAIC, BIC) depict dimensionality in a multidimensional Rasch analysis when the dichotomous data is generated by a Rasch model?

(2) How do test length, sample size, and between-dimension correlation affect the accuracy with which four global fit indices (AIC, AICc, CAIC, BIC) depict dimensionality in a multidimensional Rasch analysis of dichotomous data when the slope and lower asymptote assumptions of the Rasch model are violated?

(3) How do test length, sample size, rating scale structure and between-dimension correlation affect the accuracy with which four global fit indices (AIC, AICc, CAIC, BIC) depict dimensionality in a multidimensional Rasch analysis of polytomous data?

RATIONALE AND JUSTIFICATION FOR STUDY

Structural equation modeling (SEM) is used for confirmatory factor analysis (CFA) almost exclusively. Researchers interested in conducting an item response analysis face a two-

step process in using SEM for CFA and then scaling the data to an appropriate item response theory (IRT) model. Multidimensional item response theory could be used simultaneously for confirmatory factor analysis and scaling, but little research has been done concerning the application of global fit indices used for CFA to IRT models.

OVERVIEW OF METHODOLOGY

This project consisted of two simulation studies, one with dichotomous data and one with polytomous data. Each study had four independent variables. In both studies, test length (20 & 40 items), sample size (100, 250, 500, & 750), and between-dimension correlation (.60, .70, .80, & .90) were independent variables. Additionally, the data generation model was an independent variable for the dichotomous data, based on a one-parameter, two-parameter, and three-parameter multidimensional IRT model. The polytomous data was generated using two rating scale structures, a 3 point and a 5 point rating scale.

The data from each study was analyzed in the same manner. The simulated data was scaled to the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) using ConQuest (Wu, Adams, Wilson, & Heldane, 2007). Each simulated data set was scaled to four configurations of the MRCMLM using ConQuest. The configurations represented the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model).

Four global fit indices (AIC, AICc, CAIC, BIC) were calculated for each of the four models based the deviance statistic. One of the four models was designated as the "selected model" (i.e., the model associated with the smallest value of the global fit index) for each

simulated data set using each of the fit indices. The proportion correct and incorrect selected models for the replications within each cell of the experimental design (i.e., the dependent variable) were computed and compared between global fit indices and across experimental conditions.

A repeated measures logistic regression analysis was completed for each study to determine which effects were significant. The analysis modeled a dichotomous outcome, whether the index resulted in the correct or incorrect selection of the generating dimensionality structure, as a function of the main effects and two-way interactions of the independent variables and the indices. A repeated measures analysis was used since the same dataset was analyzed using each of the four indices. Odds ratios were used as indicators of effect sizes, with the cut-off for practical significance being an odds ratio greater than 2.0 (or less than 0.5, for negative relationships). This metric for a large effect size is common in medical sciences (Monahan, McHorney, Stump, & Perkins, 2007).

ORGANIZATION OF DISSERTATION

The first chapter of this dissertation serves as an introduction of the problem, research questions, and processes used in the two studies. The literature review is contained in the second chapter. Chapters three and four are publishable-quality articles for the dichotomous and polytomous studies, respectively. The last chapter summarizes the results from the two manuscripts, identifies limitations of the line of research, discusses the implications and impact of the research, and identifies areas for future research.

CHAPTER TWO

LITERATURE REVIEW

Confirmatory factor analysis (CFA) is frequently used to evaluate the theoretical dimensions of a measurement instrument. In CFA, it is common to compare the fit of multiple latent trait models, using one of two methods to compare the relative fit of those models: hypothesis tests, such as a chi-squared test and the chi-squared difference test, and global fit indices, including the Akaike Information Criteria (AIC; Akaike, 1973). However, dimensionality procedures in multidimensional item response theory (MIRT) have not been studied thoroughly. The purpose of this study is to compare the performance of several of the global fit indices in CFAs as conducted in applications of multidimensional item response models.

Model Selection and Global Fit Indices

Global fit indices, such as AIC, BIC (Bayesian Information Criteria; Schwarz, 1978), AICc (Bias-Corrected AIC; Hurvich & Tsai, 1989) and CAIC (Consistent AIC; Bozdogan, 1987) were designed to be used in model comparison and selection to identify the most parsimonious model, given a particular set of observations. These four particular indices are based on information theory and are frequently referred to as "penalized" model selection criteria since the formulas for these indices include a term involving the number of parameters in the model. This term serves to increase the value of the indices for models with larger number of parameters (Weakliem, 2004; Kuha, 2004). Smaller values of each index indicate the best fitting model out of the models being considered.

Although these indices are used in the same manner, there is a difference in the exact formulas (Equations 1 through 4) as well as the theory behind the indices.

$$AIC = - 2*(\log likelihood) + 2*p \quad (1)$$

$$BIC = - 2*(\log likelihood) + p*\ln(N) \quad (2)$$

$$AICc = - 2*(\log likelihood) + 2*p + (2*p*(p+1))/(N-p) \quad (3)$$

$$CAIC = -2*(\log likelihood) + p(\ln(N) + 1) \quad (4)$$

In each of the equations, *p* represents the number of parameters being estimated by the model and *N* refers to sample size.

The AIC was first proposed by Akaike in 1973 (Weakliem, 2004). AIC is considered to be an efficient criterion as it was designed around the idea that the set of models being compared does not contain the true model as the true model is unknown (McQuarrie & Tsai, 1998; Kuha, 2004). The "penalty term" in the AIC is the expected value of a chi-squared with *p* degrees of freedom (Akaike, 1987). This stems from the difference of the log likelihoods for the true and observed distributions following a chi-squared distribution.

The AIC has some known deficiencies. One problem is that that as the number of parameters increases with the sample size, the AIC becomes a biased estimator. Another problem with the AIC is that it is not a consistent statistic. As N approaches infinity, the AIC does not select the true model when that model is in the set of models being compared (Yang & Yang, 2007).

Alternative information-based criteria were developed to address these issues. AICc was developed as a small-sample bias-corrected version of AIC (McQuarrie & Tsai, 1998). As the sample size increases, AICc and AIC converge. Some have argued that AICc should be used in place of AIC unless $N/p > 40$ for the largest value of p in the models being compared, since AIC is a biased estimator when $N/p < 40$ and will overfit the model (Hurvich & Tsai, 1989; Burnham & Anderson, 2004). A consistent version of AIC (CAIC) was proposed by Bozdogen (1987).

CAIC was derived to be an asymptotically unbiased criterion for determining the true order of a model (Anderson et. al, 1998). CAIC imposes a larger penalty for over-parameterization than does AIC. The Bayesian Information Criterion (BIC), sometimes referred to as the Schwarz Information Criterion (SIC) (Schwarz, 1978), is another consistent criterion, designed under the assumption that the true model is included in the set of models being compared (McQuarrie & Tsai, 1998; Kuha, 2004). It was designed under the Bayesian framework, using an asymptotic expansion of Bayes formula.

Performance of the Information-Based Fit Indices

The majority of the research on global fit indices has compared the performance of multiple global fit indices to each other as well as other measures of fit. The focus of these studies has included various types of models: regression and/or time series models (Bozdogan, 1987; McQuarrie & Tsai, 1998, Kuha, 2004); structural equation modeling (Cudeck & Browne, 1983; Haughton et al., 1997; Marsh et al., 1988; Bandalos, 1993; Hu & Bentler, 1998; Whittaker & Stapleton, 2006); and item response models (Bolt & Lall, 2003; Kang & Cohen, 2007).

Regression

Bozdogan (1987) conducted a Monte Carlo study to compare AIC and CAIC when selecting the true degree of a polynomial regression model. Independent variables included sample size (N=50, 100, and 200) and residual error. Bozdogan found that AIC selected an over-parameterized model more often than CAIC. In addition, CAIC was more likely to select the same the same model than AIC across all of the studied sample size and residual error conditions.

McQuarrie and Tsai (1998) describe several simulation studies in which they considered sixteen fit indices for identifying appropriately specified univariate and multivariate regression

and time series models. For each simulation situation, they ranked the indices according to the percentage of time the generating model was identified. They found AICc and BIC to be superior to the AIC in terms of model identification accuracy when the true model is included in the set of considered models for small (n=15 to 100) and large sample (n=25,000) univariate regression, small sample multivariate regression, and small sample univariate autoregressive simulations. For the large sample simulations of multivariate regression, univariate autoregressive, and vector autoregressive models, AICc and AIC performed similarly in terms of accuracy, while BIC had a higher accuracy rate than either of these indices.

Kuha (2004) compared the performance of AIC and BIC in applications to two large social mobility datasets, identifying forms of log-linear models for a three-way contingency table for those data. In that application, BIC was more likely to result in the selection of less complex models compared to AIC. However, both indices had the largest values for the same models, indicating that they agreed in terms of the models that were the worst in terms of fit.

Structural Equation Modeling/Confirmatory Factor Analysis

In an early study regarding cross-validation of covariance structures, Cudeck and Browne (1983) compared the performance of a cross-validation index (CVI) to the performance of AIC and BIC in evaluating seventeen factor models. A real dataset of 2,676 observations was randomly assigned into smaller groups (N = 75, 100, 200, 400, 800, 1338) for the sake of evaluating sample size differences while when performing cross-validation. AIC was found to select more heavily parameterized models than CVI, while BIC was found to select less heavily parameterized models than CVI.

Bandalos (1993) cited the Cudeck and Browne (1983) study, mentioning that because it used real data, the true model was unknown. With this in mind, Bandolos conducted a simulation

study in which these indices were compared in the context of CFA using structural equation modeling. Sample size (N=200 and 600), factor loadings (.4, .6, .8), and model misspecifications were the independent variables. Model misspecification was created by setting secondary factor loadings to zero. Four indices, including AIC and CAIC, were compared in the study. The study revealed that AIC, BIC and CAIC were comparable in terms of selecting the true model when the sample size are greater than 600.

In another simulation study of CFA, Haughton, Oud, and Jansen (1997) compared AIC, CAIC, and BIC, along with several other indices. The independent variables were sample size (N = 100, 400, 1,000, and 6,000), model misspecification (five levels), and disparity levels among true values of parameter error variances (small, moderate and large). The study found that none of the indices performed well when N=100. Under other conditions, CAIC and BIC performed better than AIC at selecting the correctly specified model.

A study by Marsh, Balla and McDonald (1998) used real and simulated data to compare 29 fit indices, including AIC and BIC, in a CFA context by fitting a three-factor, simple structure model to four data sets. Previous analysis of the two real datasets had supported a three-factor model, with one approximating simple structure and the other approximating complex structure. For the simulated data, seven sample sizes were generated, each from a simple-structure and a complex-structure three-factor model. The study measured the independence of the performance of the indices from sample size by the extent to which the indices had the same result across the seven sample sizes for a particular dataset. The results indicated that the performance of AIC and BIC was influenced by sample size and the authors' stated that values of these global fit indices cannot be "interpreted independently of sample size".

A CFA study by Whittaker and Stapleton (2006) varied sample sizes (N=100, 200, 500, 1,000), factor loading size (.5 and .7), model misspecification, and non-normality through the introduction of for generated data. Model misspecification was by achieved through adding or removing secondary factor loadings in the true model, while non-normality was introduced through skew and kurtosis values. They evaluated each index to see how often it chose the true model as well as how consistent it was. The results showed that CAIC performs comparably to BIC and better than AIC under larger samples sizes (N = 1,000). When CAIC and BIC did not choose the true model, they consistently chose the under-parameterized model.

In summary, the previous studies in the applications of regression and structural equation modeling have shown that the use of smaller sample sizes impacts the performance of AIC, BIC, AICc, and CAIC. The studies have also shown that BIC tends to choose an under-parameterized model than AIC under certain conditions, with the accuracy of the two indices being comparable with larger sample sizes.

Item Response Theory

There have been few studies of the application of AIC, BIC, and CAIC to latent trait model selection, particularly with respect to multidimensional models. Kang and Cohen (2007) compared several indices, including AIC and BIC, when selecting between one-, two-, and three-parameter uni-dimensional item response theory (IRT) models for real and simulated data. The real data consisted of two passages from a national education exam: a dichotomously-graded multiple-choice section assumed to follow a 3-parameter logistic model (3PL) and an open-ended section that was previously determined to follow a 2-parameter logistic model (2PL). Results from this portion of the study indicated that the BIC was more likely to lead to adoption of a simpler model than the AIC, as it indicated the 2PL model for the multiple choice data while

the AIC indicated the 3PL model. Kang and Cohen also reported a simulation study in which test length (N=20, 40), sample size (N=500, 1000), ability distribution offset, and generating model (1-parameter, 2-parameter, and 3-parameter logistic models), were the independent variables. Although their analyses involved the comparison of generating and estimated model parameters, the averages of several indices, including the AIC and BIC, over the 50 replications were reported. In all cases, the average of both AIC and BIC was smallest for the generating model, indicating that they perform comparably in terms of average accuracy for larger sample sizes.

A simulation study (Harrell & Wolfe, 2009) focused on the effect of between-dimension correlation on the accuracy of four global fit indices in recovering the generating dimensional structure. Dichotomous data was generated using the Multidimensional Random Coefficient Multinomial Logit model (MRCMLM). The independent variables studied were between-dimension correlation (0, .3, .45, .6, .75, .90) and sample size (100, 250, 500, 1000). Four versions of the MRCMLM were fit to represent the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model). For each of the 100 iterations of the 24 design cells, four fit indices (AIC, BIC, AICc, CAIC) were calculated for each of the four model specifications, and model selection was conducted by comparing the values of each index across the four models. That study revealed that higher between-dimension correlations and smaller sample sizes influence the accuracy with which each index indicated the correct two-dimension generating structure. The four indices had 100% accuracy when between-dimension correlation was between .00 and .60. When between-dimension correlation was .75, accuracy was only less than optimal at

N=100. Of the four indices, when between-dimension correlation was very high (.90), AIC performed better at N=100 than the other three in terms of indicating the correct two-dimension generating structure, and at larger sample sizes AIC and AICc perform similarly and better than CAIC and BIC.

Further research in item response theory in this area should investigate the effect of other variables such as data-to-model misfit through the use of one-parameter, two-parameter and three-parameter multidimensional item response models.

Multidimensional Item Response Models

Multidimensional IRT models can be described as compensatory or non-compensatory. In a compensatory model, a low ability on one dimension can be compensated for with greater ability on another dimension. Compensatory models were the first multidimensional models developed and are the most commonly used.

The Multidimensional Random Coefficient Multinomial Logit model (MRCMLM) (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003) is a one-parameter model that depicts *D* traits underlying a person's responses to the items on the instrument, allowing item difficulty to vary, while assuming that the items have a constant slope parameter and lower asymptote of zero. These *D* latent traits are listed in a *D*x1 vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. The vector $\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, \ldots, b_{ikD})$ represents the scores across the D dimensions, where b_{ikd} is the response in category *k* in item *i* for dimension *d*. These vectors can be collected into a matrix **B** which represents the person's scoring matrix for the entire instrument.

In the case of polytomous data, item threshold parameters are represented by the vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$. The design matrix **A** represents the assumed relationship of the items to the dimensions. The design matrix consists of design vectors $\mathbf{a}_{ik}$, each of length p, for each item. The

random variable X_{ik} represents a person's response to an item. The variable takes on the value 1 if the response to item *i* equals *k*. Using the design vectors and the scoring vectors, the probability of a response of category *k* on item *i* is represented by

$\pi_{X_{ik}=1;\mathbf{A},\mathbf{B},\xi|\boldsymbol{\theta}} = \frac{\exp\left(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}{\sum_{k=1}^{K_i}\exp\left(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}$ (Adams, Wilson, & Wang, 1997). The probability can be

rewritten in model form as $\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = (\mathbf{b}'_{ik} - \mathbf{b}'_{i(k-1)})\boldsymbol{\theta}_n + (\mathbf{a}'_{ik} - \mathbf{a}'_{i(k-1)})\boldsymbol{\xi}$ (Wang, 2004).

The multidimensional two-parameter model (M2PL) (Reckase, 1985) allow for items to have different slopes and sets the lower asymptote to equal 0. The probability of a correct response to an item *i* is modeled as $P(X_i = 1 \mid \theta_j, a_i, d_i) = \frac{\exp((\sum_{k=1}^{m} a_{ik}\theta_j) + d_i)}{1 + \exp((\sum_{k=1}^{m} a_{ik}\theta_j) + d_i)}$, where θ_j = a vector of latent abilities; a_i = a vector of slope for item *i* in dimension *k*; d_i= location parameter for item *i*, related to the difficulty for item *i*.

The multidimensional compensatory three-parameter logistic model (MC3PL) (Spray, Davey, Reckase, Ackerman, & Carlson, 1990; Reckase, 1997) models item slope, difficulty, and a lower asymptote. The probability of a response of person *j* to item *i* in dimension *k* is represented by $P(X_i = 1 \mid \theta_j, a_i, d_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp(-1.7(\sum_{k=1}^{m} a_{ik}\theta_j) + d_i)}$, where θ_j = a vector of latent abilities; a_i = a vector of slope for item *i* in dimension *k*; d_i= location parameter for item *i*, related to the difficulty for item *i*; c_i = the lower asymptote parameter of item *i*.

Model Recovery with Multidimensional IRT Models

Information is lacking on the effect of highly correlated dimensions on model recovery using AIC or BIC. Bolt and Lall (2003) studied model recovery in IRT using Markov Chain Monte Carlo procedures. Their simulation study varied sample size (1000, 3000), test length (20, 40), and correlation between multidimensional ability distributions (0, .3, .6) when fitting the multidimensional two-parameter logistic model (M2PL) (Reckase, 1985, 1997) and the multidimensional latent trait model (MLTM) (Whitely, 1980). Although AIC and BIC were not used in this study, they did find that a between-dimension correlation of .60 resulted in a larger number of examinees and items being required for accurate model recovery.

The simulation study by Harrell & Wolfe (2009) investigated model recovery at higher between-dimension correlations, but their design involved large steps in the correlation values (.6, .75, .9). It is still unknown as to what happens at values in between those used in the study, including determining at which correlation value accuracy becomes a concern.

In addition, other variables that have been known to affect model recovery in IRT need to be examined. Bolt and Lall (2003) demonstrated that generating model, test length, and sample size affected model recovery in a uni-dimensional Rasch analysis when the between dimension-correlation was .60. Based on this information, it is highly probable that these variables affect the accuracy of the global fit indices in multidimensional Rasch analysis as well. The effect of these variables on the analysis of polytomous data has not been studied. Since dichotomous data is polytomous data with only two categories, it is highly probable that moderate to high between-dimension correlation should affect the performance of the fit indices in an analysis of polytomous data.

CHAPTER THREE

MANUSCRIPT 1: THE INFLUENCE OF MISFIT ON MULTIDIMENSIONAL RASCH MODEL GLOBAL FIT INDEX ACCURACY

Abstract

Most research on confirmatory factor analysis using global fit indices (AIC, BIC, AICc, and CAIC) has been in the structural equation modeling framework. Little research has been done concerning application of these indices to item response models, especially within the framework of multidimensional Rasch analysis. The results of a simulations study that investigated how sample size, between-dimension correlation, model-to-data misfit, and test length affect the accuracy of these indices in model recovery in dichotomous data using a multidimensional Rasch analysis are described in this dissertation. The interaction effect between global fit index and between-dimension correlation had a very large effect size. At higher values of between-dimension correlation, AIC indicated the correct two-dimension generating structure slightly more often than does the BIC or CAIC. The correlation by index interaction also had an odds ratio indicating practical importance. As expected, the results demonstrated that violations of the Rasch model assumptions are magnified at higher between-dimension correlations. Recommendations for practitioners working with highly correlated multidimensional data include creating moderate length (roughly 40 items) instruments, minimizing data-to-model misfit in the choice of model used for confirmatory factor analysis (MRCMLM or other MIRT models), and making decisions based on multiple global indices instead of depending on one index in particular.

Introduction

Confirmatory factor analysis (CFA) is frequently used to evaluate the theoretical dimensions of a measurement instrument. Both the number of factors and the pattern of how the items correspond to the factors (factor loadings) are verified through CFA. The results of such analyses serve as structural validity evidence for scores based on responses to the instrument. Currently, CFA is conducted almost exclusively within an SEM framework. This makes latent trait model scaling of measures from an instrument a two step-process: conducting CFA using software such as Lisrel or MPlus and then scaling the verified structure to a latent trait model (e.g., the Rasch model) using software such as Winsteps or Conquest. With the development of multidimensional item response (MIRT) models, it is possible to conduct confirmatory factor analysis within a latent trait test theory framework.

In CFA, it is common to compare the fit of multiple theory-driven latent trait models, using one of two methods to compare the relative fit of those models: hypothesis tests, such as a chi-squared-based likelihood ratio test, and global fit indices, including the Akaike Information Criteria (AIC; Akaike, 1973; Akaike, 1987). The chi-squared test is known to be sensitive to sample size which may result in statistical significance even when the differences between the estimated and observed covariance matrices are negligible (Brown, 2006). Hence, the use of fit indices is prevalent in CFA. Global fit indices, such as AIC, BIC (Bayesian Information Criteria; Schwarz, 1978), AICc (Bias-Corrected AIC; Hurvich & Tsai, 1989) and CAIC (Consistent AIC; Bozdogan, 1987) were designed to be used in model comparison and selection to identify the most parsimonious model, given a particular set of observations. These four particular indices are based on information theory and are frequently referred to as "penalized" model selection criteria because the formulas for these indices include a term involving the number of parameters

in the model. This term serves to increase the value of the indices for models with larger numbers of parameters (Weakliem, 2004; Kuha, 2004). Smaller values of each index indicate a better fitting model.

Although these indices are used in the same manner, their formulas differ (Equations 1 through 4).

$$AIC = - 2*(\log likelihood) + 2*p \quad (1)$$

$$BIC = - 2*(\log likelihood) + p*\ln(N) \quad (2)$$

$$AICc = - 2*(\log likelihood) + 2*p + (2*p*(p+1))/(N-p) \quad (3)$$

$$CAIC = -2*(\log likelihood) + p*\ln(N) + p \quad (4)$$

In each of the equations, *p* represents the number of parameters being estimated by the model and *N* refers to sample size. Since AIC has the smallest penalty, it will have the smallest value of the four indices. It is likely to select a different model than the other three indices, one that is more complex. BIC and CAIC should result in very similar numbers, but CAIC will be greater and will increase at a quicker rate than BIC as the number of parameters increases. This suggests that CAIC should be more likely to result in selection of simpler models than BIC. AICc has values that are between those of BIC and AIC. As sample size increases, the value of AICc decreases and it approaches the value of AIC. Hence, AICc is more likely to indicate the same model as AIC.

Some of the fit indices used in SEM, such as AIC and BIC, may be utilized in confirmatory factor analysis using multidimensional item response models. However, there is little research on the performance of these fit indices for the sake of conducting a CFA using this framework. This study seeks to examine the accuracy of the global fit indices (AIC, AICc, CAIC, and BIC) in a multidimensional Rasch analysis for dichotomously-scored data generated

from multi-dimensional item response models, under conditions of short and moderate length tests, small to moderate sample sizes, and moderate to high correlation between dimensions on a simulated two-dimensional instrument. The study also examines the effects of model-to-data misfit on the accuracy of global fit indices as applied to CFA within an MIRT framework.

Research on Global Fit Indices

The majority of the research on global fit indices has compared the performance of multiple global fit indices to each other. The focus of these studies has included regression and/or time series models (Bozdogan, 1987; McQuarrie & Tsai, 1998, Kuha, 2004) and structural equation modeling (Cudeck & Browne, 1983; Haughton et al., 1997; Marsh et al., 1988; Bandalos, 1993; Hu & Bentler, 1998; Whittaker & Stapleton, 2006). The results have indicated that the use of smaller sample sizes impacts the performance of AIC, BIC, AICc, and CAIC. The studies have also shown that BIC tends to specify under-parameterized models as being the best fitting more often than AIC under certain conditions, with the accuracy of the two indices being comparable with larger sample sizes.

There have been few studies of the application of AIC, BIC, and CAIC to latent trait model selection, particularly with respect to multidimensional models. Kang and Cohen (2007) compared several indices, including AIC and BIC, when selecting between one-, two-, and three-parameter uni-dimensional item response theory (IRT) models for real and simulated data. The real data consisted of two passages from a national education exam: a dichotomously-graded multiple-choice section assumed to follow a 3-parameter logistic model (3PL) and an open-ended section that was assumed to follow a 2-parameter logistic model (2PL). Results from this portion of the study suggest that the BIC may be more likely to lead to adoption of a simpler model than the AIC, as it indicated the 2PL model for the multiple choice data while the AIC

indicated the 3PL model. Kang and Cohen also reported a simulation study in which test length (N=20, 40), sample size (N=500, 1000), ability distribution offset, and generating model (1-parameter, 2-parameter, and 3-parameter logistic models), were the independent variables. Although their analyses involved comparison of generating and estimated model parameters, the article reports the averages of several indices, including the AIC and BIC, over the 50 replications of the simulation. In all cases, the average of both AIC and BIC was smallest for the generating model, indicating that they may perform comparably in terms of average accuracy for larger sample sizes.

A simulation study (Harrell & Wolfe, 2009) focused on the effect of between-dimension correlation on the accuracy of four global fit indices in identifying generating dimensional structures. In that study, dichotomous data was generated using the Multidimensional Random Coefficient Multinomial Logit model (MRCMLM) (Adams, Wilson & Wang, 1997). The independent variables were between-dimension correlation (.00, .30, .45, .60, .75, .90) and sample size (100, 250, 500, 1000). Four versions of the MRCMLM were fit to represent the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model). For each of the 100 iterations of the 24 experimental design cells, four fit indices (AIC, BIC, AICc, CAIC) were calculated for each of the four model specifications, and model selection was conducted by comparing the values of each index across the four models. That study revealed that higher between-dimension correlations and smaller sample sizes influence the accuracy with which each index indicated the correct two-dimension generating structure. The four indices had 100% accuracy when between-

dimension correlation was between .00 and .60. When between-dimension correlation was .75, accuracy was only less than optimal at N=100. Of the four indices, when between-dimension correlation was very high (.90), AIC performed better at N=100 than the other three in terms of indicating the correct two-dimension generating structure, and at larger sample sizes AIC and AICc perform similarly and better than CAIC and BIC.

The current study builds on the work of Kang & Cohen (2007) and Harrell & Wolfe (2009) to further investigate the use of the global fit indices as applied to CFA in an item response theory framework. Our focus is on the accuracy of decisions in a multidimensional Rasch analysis of dichotomous data using four global fit indices as a function of sample size, number of test items, data-to-model misfit, and between-dimension correlation. It is hypothesized that smaller sample sizes, shorter tests, higher levels of between-dimension correlation will result in decreases in accuracy. It is also anticipated that higher levels of data-to-model misfit will decrease the accuracy of the indices.

Methods

Simulation Design

Data was simulated using SAS and Multisim to represent dichotomous two-dimensional item response data. Examinee ability was generated from bivariate standard normal distribution. Item difficulty generating parameters were drawn from a standard normal distribution. Item slope parameters (for two- and three- parameter models) were drawn from a lognormal(0,.25) distribution. Lower asymptote parameters (for the three-parameter model) were set to a constant value of .20.

There were four independent variables in this study. Between-dimension ***correlation*** had moderate to high values (.60, .70, .80, & .90). These values were chosen based on Harrell and

Wolfe (2009), which found that the accuracy of the fit indices was not affected at values of .60 or lower. ***Sample size*** (100, 250, 500, & 750) represented smaller, moderate and larger sample sizes. ***Test length*** (20 & 40) modeled shorter and moderate length tests. Based on results in Kang and Cohen (2007), the data was simulated from multidimensional Rasch (MRCMLM), 2 parameter (M2PL) and 3 parameter (MC3PL) IRT models (***Generating model***). There were 200 replications for each cell of the 96 cell [4 (correlation) x 4 (sample size) x 2 (test length) x 3 (generating model)] experimental design. In all of the cases, two dimensions were simulated with half of the items defining each dimension.

Analysis

Each simulated data set was scaled to four configurations of the MRCMLM using ConQuest (Wu, Adams, Wilson, & Heldane, 2007). The configurations represented the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model). Table 3.1 represents the mappings of the items for each of these model specifications for two test lengths.

Table 3.1. Item-to-dimension mappings

Test Length	Specification	Number of Dimensions	Items in Dimension 1	Items in Dimension 2	Items in Dimension 3
20 Items	Underfit	1	1-20		
	Correct	2	1-10	11-20	
	Misspecified	2	1-5, 16-20	6-15	
	Overfit	3	1-5	5-10	11-20
40 Items	Underfit	1	1-40		
	Correct	2	1-20	21-40	
	Misspecified	2	1-10, 31-40	11-30	
	Overfit	3	1-20	21-30	31-40

The values of AIC, AICc, BIC, and CAIC were calculated for each scaled data set using the log likelihood values produced by ConQuest. The minimum value of each fit index was used as the selection criterion. The model satisfying the criterion for the index in question was designated as the "selected model" for each index for each simulated data set. The proportion of data sets for the 200 replications within each cell of the experimental design that selected each of the 4 model specifications (i.e., the dependent variable) was computed and compared between experimental conditions.

A repeated measures logistic regression analysis was completed for each study to determine which effects were significant. The analysis modeled a dichotomous outcome, whether the index resulted in the correct or incorrect selection of the generating dimensionality structure, as a function of the main effects and two-way interactions of the independent variables and the indices. A repeated measures analysis was used since the same dataset was analyzed using each of the four indices. Odds ratios were used as indicators of effect sizes, with the cut-off for practical significance being an odds ratio greater than 2.0 (or less than 0.5, for negative relationships). This metric for a large effect size is common in medical sciences (Monahan, McHorney, Stump, & Perkins, 2007).

Results

Table 3.2 presents the results of a logistic regression analysis that contains two-way interaction terms crossing each of the independent variables with the global fit indices. Two of the interaction effects had large effect sizes (i.e. odds ratios greater than 2 or less than .50): "correlation x index" and "correlation x generating model".

Table 3.2. Hypothesis Testing and Effect Sizes from Logistic Regression

Effect	Test Statistic	p-value	Odds Ratio
Main Effect			
Sample Size (N)	4.99	< .0001	1.01
Test Length (NI)	32.32	< .0001	1.60
Between-Dim. Correlation (CORR)	- 5.67	< .0001	0.02
MIRT Generating Model: M1PL	29.27	< .0001	135.59
MIRT Generating Model: M2PL	25.41	< .0001	525.47
Index: AIC	-5.71	< .0001	.46
Index: BIC	1.89	.0583	1.05
Index: AICc	-6.18	< .0001	0.59
Interaction Effects			
N*AIC	-495.50	< .0001	1.00
N*AICC	-304.22	< .0001	1.00
N*BIC	-624.62	< .0001	1.00
NI*AIC	-40.68	< .0001	0.97
NI*AICC	-26.90	< .0001	0.97
NI*BIC	-33.70	< .0001	0.99
CORR*AIC	29.25	< .0001	34.47
CORR*AICC	34.04	< .0001	18.5
CORR*BIC	30.19	< .0001	2.06
NI*CORR	-24.03	< .0001	0.65
NI*M1PL	23.12	< .0001	1.03
NI*M2PL	1.18	.2396	1.00
N*CORR	-3.60	.0003	0.99
N*NI	1.48	.1393	1.00
N*M1PL	-5.78	< .0001	1.00
N*M2PL	-7.78	< .0001	1.00
CORR*M1PL	-29.72	< .0001	0.03
CORR*M2PL	-16.55	< .0001	0.01
AIC*M1PL	-68.94	< .0001	0.50
AICC*M1PL	-69.15	< .0001	0.61
BIC*M1PL	-56.52	< .0001	0.88
AIC*M2PL	-197.69	< .0001	0.57
AICC*M2PL	-170.98	< .0001	0.49
BIC*M2PL	-91.64	< .0001	0.86

Note: The reference levels for the categorical variables in the model are not listed: MC3PL generating model, the index CAIC, and the interactions that include these terms.

The odds ratios for the interaction between correlation and index suggest that the indices perform differently based on the strength of the between-dimension correlation. The interaction plot in Figure 3.1 shows that CAIC performs worse than AIC at a correlation of .90 while all

indices performed comparably at correlations less than or equal to .80. While accuracy is almost perfect at a correlation of .60, Table 3.3 reveals that all of the indices perform at below-optimal levels once correlation is greater than .70.

Figure 3.1. Correlation by Index Interaction Plot

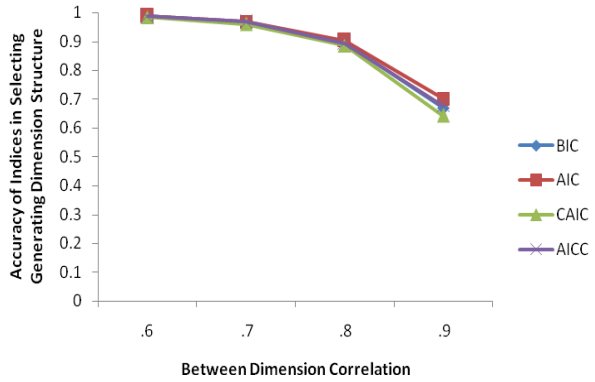

Table 3.3. Accuracy Results for Indices by Correlation for Between Dimension Correlation ≥ .70

	Between-Dimension Correlation											
	.70				.80				.90			
Index	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3
AIC	**	.97	**	.02	.02	.90	.01	.07	.06	.70	.05	.19
BIC	.02	.97	**	**	.06	.90	.01	.03	.19	.67	.06	.08
AICc	.01	.97	**	.01	.04	.89	.06	.01	.12	.68	.04	.16
CAIC	.03	.96	**	**	.08	.89	.01	.02	.24	.65	.05	.06

Table 3.3 also reveals which dimension structure the indices select when they do not indicate the generating dimension structure at the higher levels of correlation. At correlations of .80 and .90, AIC is about three times as likely to indicate the three-dimension model as it is the one-dimension structure. BIC is at least twice as likely to indicate the one-dimension model as the three dimension model at correlations of .80 and .90. At the .80 level of correlation, AICc is four times as likely to indicate the one-dimension model as the three-dimension model. However, at the .90 level, AICc is almost equally likely to select the one-dimension or three-

dimension model. CAIC is almost four times as likely to indicate the one-dimension model as it is three dimension model.

The correlation by generating model interactions had very large effect sizes, indicating that the effect of the level of between-dimension correlation is affected by the generating model (M1PL, M2PL, M3PL). Figure 3.2 displays the accuracy results in terms of correlation levels for each generating model. None of the indices performed well for data generated from the MC3PL model at the three higher correlations (.70, .80, & .90) when compared to the performance for data generated from the M1PL and M2PL models. Table 3.4 reveals which model is being indicated when the correct dimensional structure is not indicated by the indices. For the M1PL- and M2PL-generated data, the selection pattern is similar, with incorrect selections almost equally distributed between the one dimension and three dimension model. A different pattern is revealed for the M3PL-generated data, where the one-dimension model is selected almost twice as often as the three-dimension model.

Figure 3.2. Correlation by Generating Model Interaction Plot

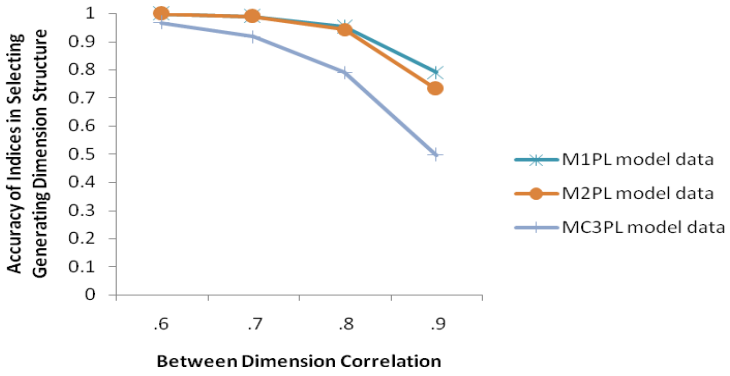

Table 3.4. Accuracy Results by Generating Model and Between-Dimension Correlation (in %)

	One-Parameter MIRT Model (M1PL)				Two-Parameter MIRT Model (M2PL)				Three-Parameter MIRT Model (MC3PL)			
r	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3
.6	0	.99	0	**	**	.99	0	**	.02	.96	**	.01
.7	**	.99	**	**	**	.99	**	**	.04	.92	**	.03
.8	.01	.95	**	.03	.02	.94	**	.03	.12	.79	.02	.07
.9	.08	.79	.03	.10	.09	.73	.05	.13	.28	.50	.08	.14

Note: r= Between-Dimension Correlation; 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure; ** indicates that the dimensional structure was indicated less than 1% of the time.

For the sake of completeness, tables 3.5 through 3.7 show the results for index in most of the cells of the design for this study. The tables are organized by generating model and correlation due to the large effect size of this interaction. Results for the between-dimension correlation of .60 are not shown since the indices had perfect or near perfect accuracy at this level.

Table 3.5. Accuracy Results for Data Generated by One-Parameter MIRT Model for Between-Dimension Correlation ≥ .7 (in %)

			Model Selection Methods															
			AIC				BIC				AICc				CAIC			
r	NI	N	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3
. 7	20	100	0	95	.5	4.5	3	95.5	0	1.5	3	95.5	0	1.5	7.5	91.5	.5	.5
		250	0	98	0	2	0	99	0	1	0	99	0	1	0	100	0	0
		500	0	98.5	0	1.5	0	99	0	1	0	99	0	1	0	99	0	1
		750	0	99.5	0	.5	0	100	0	0	0	100	0	0	0	100	0	0
	40	100	0	99	0	1	0	99	0	1	0	99	0	1	0	99	0	1
		250	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
		500	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
		750	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
.8	20	100	1	92	1	6	11	86.5	.5	2	3.5	93	1	2.5	19	79.5	.5	1
		250	0	90	.5	9.5	1	92.5	2	4.5	0	91.5	.5	8	1	93.5	2	3.5
		500	0	90.5	0	9.5	0	94.5	0	5.5	0	91.5	0	8.5	0	96.5	.5	3
		750	0	90	0	10	0	93.5	0	6.5	0	90	0	10	0	94.5	0	5.5
	40	100	0	96.5	0	3.5	0	99	.5	.5	.5	99	.5	0	0	99.5	.5	0
		250	0	98.5	0	1.5	0	98.5	0	1.5	0	98.5	0	1.5	0	98.5	0	1.5
		500	0	99.5	0	.5	0	99.5	0	.5	0	99.5	0	.5	0	99.5	0	.5
		750	0	99.5	0	.5	0	100	0	0	0	99.5	0	.5	0	100	0	0
.9	20	100	16	55	16	13	46.5	42	10	1.5	29.5	54.5	13.5	2.5	59	34	6.5	.5
		250	2	74.5	4.5	19	19	66.5	8.5	6	4	74.5	4	17.5	25.5	62	8.5	4
		500	0	68.5	1	30.5	3.5	75	3.5	18	0	69	1	30	6	75	3.5	15.5
		750	0	69	0	31	.5	75.5	3	21	0	69	0	31	.5	76.5	3.5	19.5
	40	100	0	91	3	6	6.5	90	2.5	1	13.5	84	2	.5	15	82.5	2	.5
		250	0	95.5	0	4.5	.5	97.5	1	1	0	96	0	4	.5	1	97.5	1
		500	0	95.5	0	4.5	0	96.5	0	3.5	0	96	0	4	0	97	0	3
		750	0	92	0	8	0	93	0	7	0	92	0	8	0	94	0	6

Note: r= Between-Dimension Correlation; NI = Test Length; N = Sample Size; 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure

Table 3.6. Accuracy Results for Data Generated by Two-Parameter MIRT Model for Between-Dimension Correlation ≥ .7 (in %)

			Model Selection Methods															
			AIC				BIC				AICc				CAIC			
R	NI	N	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3
. 7	20	100	.5	94.7	1.1	3.7	2.1	96.3	1.1	.5	1.1	96.7	1.1	1.1	3.7	94.7	1.1	.5
		250	0	96.9	0	3.1	0	99.5	0	.5	0	97.5	0	2.5	0	99.5	0	.5
		500	0	98	0	2	0	98.5	0	1.5	0	98	0	2	0	99	0	1
		750	0	98.5	0	1.5	0	99	0	1	0	98.5	0	1.5	0	99	0	1
	40	100	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
		250	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
		500	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
		750	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
.8	20	100	2.6	89	3.1	5.3	14.1	81.7	3.7	.5	8.4	86.9	4.2	.5	22.5	74.9	2.1	.5
		250	.5	86.2	.5	12.8	1	89.3	5.1	4.6	.5	87.3	1.5	10.7	1	90.8	5.1	3.1
		500	0	90	0	10	0	96	.5	3.5	0	91	0	9	0	97	.5	2.5
		750	0	88.9	0	11.1	0	93	.5	6.5	0	88.9	0	11.1	0	94	.5	5.5
	40	100	0	98.8	0	1.1	.5	99.5	0	0	.5	99.5	0	0	.5	99.5	0	0
		250	0	98.5	0	1.5	0	98.5	0	1.5	0	98.5	0	1.5	0	99.5	0	.5
		500	0	99.5	0	.5	0	100	0	0	0	99.5	0	.5	0	100	0	0
		750	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
.9	20	100	17.3	59.1	11	12.6	52.4	37.7	7.3	2.6	37.7	46.6	10	3.7	66.5	28.3	4.2	1
		250	1.5	69.2	6.1	23.2	23.7	59.6	11.6	5.1	3.5	69.7	6.6	20.2	30.8	54	11.6	3.6
		500	0	62	2	36	4	71.5	9	15.5	0	62.5	2	35.5	6	70.5	10	13.5
		750	0	66.3	0	33.7	0	75.4	5	19.6	0	66.3	1	32.7	0	77.4	6	16.6
	40	100	1.6	79.6	7.5	11.3	11.3	76.3	10.8	1.6	18.8	72.6	8.6	0	16.1	73.7	10.2	0
		250	0	86.1	1.5	12.4	.5	92.8	3.6	3.1	0	89.2	2.6	8.2	1	92.3	4.1	2.6
		500	0	83.8	0	16.2	0	89.3	1.5	9.1	0	83.8	0	16.2	0	91.4	1.5	7.1
		750	0	85.9	0	14.1	0	90.9	0	9.1	0	85.9	0	14.1	0	91.4	0	8.6

Note: r= Between-Dimension Correlation; NI = Test Length; N = Sample Size; 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure

Table 3.7. Accuracy Results for Data Generated by Three-Parameter MIRT Model for Between-Dimension Correlation ≥ .7 (in %)

			Model Selection Methods															
			AIC				BIC				AICc				CAIC			
r	NI	N	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3	1	2C	2M	3
. 7	20	100	9.3	77.7	5.2	7.8	39.4	55.4	3.6	1.6	52.3	45.1	2.1	.5	23.3	69.4	5.2	2.1
		250	0	88.9	1	10.1	7.1	89.4	0	3.5	0	89.9	1	9.1	9.1	87.9	0	3
		500	0	91	0	9	0	94.5	0	5.5	0	92	0	8	0	95	.5	4.5
		750	0	91.5	0	8.5	0	96	0	4	0	92.5	0	7.5	0	96	0	4
	40	100	0	95.8	.5	3.7	1.6	96.9	.5	1	1.6	97.4	.5	.5	1.6	97.4	.5	.5
		250	0	99	0	1	0	99.5	0	.5	0	99	0	1	0	99.5	0	.5
		500	0	99	0	1	0	100	0	0	0	100	0	0	0	100	0	0
		750	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
.8	20	100	24.5	58.3	12	5.2	65.1	29.7	4.7	.5	47.4	43.7	7.3	1.6	0	73.4	22.9	3.7
		250	4.6	72.6	3	19.8	29.4	62.9	4.1	3.6	5.6	73.6	5.6	15.2	39.1	54.3	5.6	1
		500	0	73.5	.5	26	8	78.5	3	10.5	0	73.5	.5	26	11.5	77	3	8.5
		750	0	76	0	24	1.5	82.5	2	14	0	76	0	24	2	84.5	2	11.5
	40	100	3.6	89.1	3.1	4.2	12.5	84.4	3.1	0	24	74.5	1.5	0	19.3	78.1	2.6	0
		250	0	95.5	0	4.5	0	97	1	2	0	95.5	0	4.5	.5	97	1	1.5
		500	0	98.5	0	1.5	.5	98	.5	1	0	98.5	0	1.5	.5	98	.5	1
		750	0	96	0	4	0	97	0	3	0	96	0	4	0	97.5	.5	2
.9	20	100	44	29.6	18.1	8.3	84.5	10.9	4.1	.5	71	17.6	9.8	1.6	91.2	6.7	2.1	0
		250	17.7	40	16.7	25.8	65.2	19.2	13.1	2.5	23.2	40.9	17.2	18.7	72.7	14.7	11.1	1.5
		500	6	48	6	40	35.5	40	13	11.5	6	49.5	7	37.5	43	36	12	9
		750	1.5	47.5	3.5	47.5	25	48	7.5	19.5	2	47.5	3.5	47	31.5	44.5	8	16
	40	100	25.1	53	12.8	9.1	56.7	36.9	5.9	.5	70.6	25.7	3.2	.5	67.4	28.4	3.7	.5
		250	1.5	79.1	5.1	14.3	30.1	61.7	4.1	4.1	4.1	79.6	6.6	9.7	35.2	57.7	3.6	3.6
		500	0	75.9	5	19.1	4.5	77.4	9.1	9	0	76.4	5	18.6	9.6	74.4	8	8
		750	0	77.5	2.5	20	1.5	82.5	7.5	8.5	0	77.5	3	19.5	1.5	82.5	7.5	8.5

Note: r= Between-Dimension Correlation; NI = Test Length; N = Sample Size; 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure

Conclusions

The interaction effects between global fit index and both correlation and data-to-model misfit had very large effect sizes. While accuracy sharply decreased at a correlation of .90 for the data generated by the one- and two-parameter models, accuracy sharply decreased at a correlation of .70 for the data generated by the three-parameter model. This finding is consistent with prior research that revealed that higher levels of between-dimension correlation negatively impact the accuracy of dimensionality decisions in applications of dimensionality analysis in MIRT contexts (Harrell & Wolfe, 2009). In addition, these results demonstrate that violations of the Rasch model assumptions are magnified at higher between-dimension correlations. If data exhibits such violations, adoption of the MRCMLM may lead to incorrect decisions concerning the dimensionality of the data because these global fit indices are sensitive to both departures in the data from the modeled item structure as well as from the modeled dimensional structure. Depending on the nature of the violation, scaling the data to the M2PL or MC3PL may be more appropriate.

Higher values of between-dimension correlation lead BIC and CAIC to indicate the correct two-dimension generating structure slightly less often than does the AIC. However, the observed differences may not have practical significance. In fact, with correlations greater than .70, none of the indices perform optimally (at most 5% error). The indices differ in terms of what dimensional structure they indicate when the decision is inaccurate. BIC and CAIC tends to indicate the one-dimension model, while AIC indicates the three-dimension model when the incorrect model is chosen. AICc is equally likely to select the one-dimension or three-dimension model. In reality, researchers are likely to model two dimensions with correlation of .90 as being unidimensional, which could be misleading based on the results here. In addition, previous

research by Harrell and Wolfe (2009) suggests that when an index indicated the incorrect model, the one-dimension model was typically chosen, with the three-dimension model chosen at most 4% of the time. However, this discrepancy is most likely explained by the difference in the item-to-dimension mapping for the three-dimension structure in these studies.

Another trend existed in the descriptive statistics for these data, although it was not fully explored in the results of the logistic regression analysis. The main effect for sample size and the interaction effect for sample size with index did not have large effect sizes. However, the Harrell and Wolfe (2009) study indicated that the accuracy of the indices was sometimes questionable even at a sample size of 250. This would be reason for concern, since the rule of thumb for confirmatory factor analysis is to use at least a sample size of 200.

Further research is also warranted in terms of the way that Rasch-violations are modeled in simulation studies such as this. In our study, the two and three-parameter models were generated with a reasonable amount of variability in the slope violation. Further research simulating a narrower range of slopes should be investigated to further explore this finding. In addition, our results can be extended by examining the accuracy of dimensionality decisions based upon global fit indices when data are scaled to the M2PL and MC3PL.

In summary, the findings in this study show that the accuracy of the indices is affected by higher levels of between-dimension correlation and that the assumption of unidimensionality in such cases is not always correct based on these global fit indices. Additionally, data-to-model misfit in the presence of higher levels of correlation also affects the accuracy with which the indices select the generating dimension structure. Caution should be used when conducting confirmatory factor analysis with goodness of fit indices, such as AIC and BIC, under such conditions.

References

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov&F. Csaki (Eds.), Second international symposium on information theory (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317–332.

Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28,* 351–374.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Harrell, L. M., & Wolfe, E.W. (2009). *Effect of between-dimension correlation and sample size on multidimensional Rasch analysis.* Paper to be presented at annual meeting of the American Educational Research Association, San Diego, CA.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Communication in Statistics. Part B: Simulation & Computation, 26,* 1477–1516.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.

Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika,* 76 (2), 297-307.

Kang, T. & Cohen, A.S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement,* 31 (4), 331-358.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research,* 33 (2), 188-229.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

McQuarrie, A. D. R. & Tsai, C.L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

Monahan, P. O., McHorney, C. A., Stump, T.E., & Perkins, A.J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics,* 32 (1), 92-109.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6 (2), 461-464.

Weakliem, D.L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research,* 33 (2), 167-187.

Whittaker, T. A. & Stapleton, L. M. (2006). The performance of cross-validation indices used to select competing covariance structure models under multivariate nonnormality conditions. *Multivariate Behavioral Research,* 41 (3), 295-335.

Wu, M. L., Adams, R. J., Wilson, M. R., & Heldane, S.A. (2007). ACER ConQuest: Generalized item response modeling software (Version 2.0) [computer software]. Melbourne: Australian Council for Educational Research.

CHAPTER 4

MANUSCRIPT 2: THE ACCURACY OF GLOBAL MODEL FIT INDICES AS INDICATORS OF MULTIDIMENSIONALITY IN POLYTOMOUS DATA

Abstract

Most research on confirmatory factor analysis using global fit indices (AIC, BIC, AICc, and CAIC) has been in the structural equation modeling framework. Little research has been done concerning application of these indices to item response models, especially within the framework of multidimensional Rasch analysis. The results of a simulation study that investigated how sample size, between-dimension correlation, rating scale structure, and test length affect the accuracy of these indices in model recovery in polytomous data using a multidimensional Rasch analysis are described here. The interaction effect between global fit index and between-dimension correlation had very large effect size. At higher values of between-dimension correlation, AIC indicated the correct two-dimension generating structure slightly more often than does the BIC or CAIC. The correlation by test length interaction had an odds ratio indicating practical importance. The combination of shorter tests and higher correlations resulted in a difficult-to-detect distinction being modeled with less statistical information. Recommendations for practitioners working with highly correlated multidimensional data include creating moderate length (roughly 40 items) instruments and making decisions based on multiple global indices instead of depending on one index in particular.

Introduction

Confirmatory factor analysis (CFA) is frequently used to evaluate the theoretical dimensions of a measurement instrument. Both the number of factors and the pattern of how the items correspond to the factors (factor loadings) are verified through CFA. The results of such

analyses serve as structural validity evidence for scores based on responses to the instrument. Currently, CFA is conducted almost exclusively within an SEM framework. This makes latent trait model scaling of measures from an instrument a two step-process: conducting CFA using software such as Lisrel or MPlus and then scaling the verified structure to a latent trait model (e.g., the Rasch model) using software such as Winsteps or Conquest. With the development of multidimensional item response (MIRT) models, it is possible to conduct confirmatory factor analysis within a latent trait test theory framework.

In CFA, it is common to compare the fit of multiple theory-driven latent trait models, using one of two methods to compare the relative fit of those models: hypothesis tests, such as a chi-squared-based likelihood ratio test, or via global fit indices, such as the Akaike Information Criteria (AIC; Akaike, 1973, 1987). The chi-squared test is known to be sensitive to sample size which may result in statistical significance even when the differences between the estimated and observed covariance matrices are negligible (Brown, 2006). Hence, the use of fit indices is prevalent in CFA. Global fit indices, such as AIC, BIC (Bayesian Information Criteria; Schwarz, 1978), AICc (Bias-Corrected AIC; Hurvich & Tsai, 1989) and CAIC (Consistent AIC; Bozdogan, 1987) were designed to be used in model comparison and selection to identify the most parsimonious model, given a particular set of observations. These four indices are based on information theory and are frequently referred to as "penalized" model selection criteria because the formulas for these indices include the number of parameters in the model. This term serves to increase the value of the indices for models with greater numbers of parameters (Weakliem, 2004; Kuha, 2004). Smaller values of each index indicate a better fitting model.

Although these indices are used in the same manner, their formulas differ (Equations 1 through 4).

$$AIC = - 2*(\log likelihood) + 2*p \quad (1)$$

$$BIC = - 2*(\log likelihood) + p*\ln(N) \quad (2)$$

$$AICc = - 2*(\log likelihood) + 2*p + (2*p*(p+1))/(N-p) \quad (3)$$

$$CAIC = -2*(\log likelihood) + p*(\ln(N)) + p \quad (4)$$

In each of the equations, *p* represents the number of parameters being estimated by the model and *N* refers to sample size. Since AIC has the smallest penalty, it will have the smallest value of the four indices. It is likely to select a different model than the other three indices, one that is more complex. BIC and CAIC should result in very similar numbers, but CAIC will be greater and will increase at a quicker rate than BIC as the number of parameters increases. This suggests that CAIC should be more likely to result in selection of simpler models than BIC. AICc has values that are between those of BIC and AIC. As sample size increases, the value of AICc decreases and it approaches the value of AIC. Hence, AICc is more likely to indicate the same model as AIC.

Some of the fit indices used in SEM, such as AIC and BIC, may be utilized in confirmatory factor analysis using multidimensional item response models. However, there is little research on the performance of these fit indices for the sake of conducting a CFA within this framework. This study seeks to examine the accuracy of the global fit indices (AIC, AICc, CAIC, and BIC) in a multidimensional Rasch analysis for polytomously-scored data generated from multi-dimensional item response models, under conditions of short and moderate length tests, small to moderate sample sizes, and moderate to high correlation between dimensions on a simulated two-dimensional instrument. The study also examines the effects of model-to-data misfit on the accuracy of global fit indices as applied to CFA within an MIRT framework.

Research on Global Fit Indices

The majority of the research on global fit indices has compared the performance of multiple global fit indices to each other. The focus of these studies has included regression and/or time series models (Bozdogan, 1987; McQuarrie & Tsai, 1998, Kuha, 2004) and structural equation modeling (Cudeck & Browne, 1983; Haughton , Oud, & Jansen, 1997; Marsh, Balla, & McDonald, 1988; Bandalos, 1993; Hu & Bentler, 1998; Whittaker & Stapleton, 2006). The results have indicated that the use of smaller sample sizes impacts the performance of AIC, BIC, AICc, and CAIC. The studies have also shown that BIC tends to specify under-parameterized models as being the best fitting more often than AIC under certain conditions, with the accuracy of the two indices being comparable with larger sample sizes.

There have been few studies of the application of AIC, BIC, and CAIC to latent trait model selection, particularly with respect to multidimensional models. Kang and Cohen (2007) compared several indices, including AIC and BIC, when selecting between one-, two-, and three-parameter uni-dimensional item response theory (IRT) models for real and simulated data. The real data consisted of two passages from a national education exam: a dichotomously-graded multiple-choice section assumed to follow a 3-parameter logistic model (3PL) and an open-ended section that was assumed to follow a 2-parameter logistic model (2PL). Results from this portion of the study suggest that the BIC may be more likely to lead to adoption of a simpler model than the AIC, as it indicated the 2PL model for the multiple choice data while the AIC indicated the 3PL model. Kang and Cohen also reported a simulation study in which test length (N=20, 40), sample size (N=500, 1000), ability distribution offset, and generating model (1-parameter, 2-parameter, and 3-parameter logistic models), were the independent variables. Although their analyses involved comparison of generating and estimated model parameters, the

article reports the averages of several indices, including the AIC and BIC, over the 50 replications of the simulation. In all cases, the average of both AIC and BIC was smallest for the generating model, indicating that they may perform comparably in terms of average accuracy for larger sample sizes.

Another simulation study (Harrell & Wolfe, 2009a) focused on the effect of between-dimension correlation on the accuracy of four global fit indices in identifying generating dimensional structures. In that study, dichotomous data was generated using the Multidimensional Random Coefficient Multinomial Logit model (MRCMLM) (Adams, Wilson, & Wang, 1997). The independent variables were between-dimension correlation (.00, .30, .45, .60, .75, .90) and sample size (100, 250, 500, 1000). Four versions of the MRCMLM were fit to represent the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model). For each of the 100 iterations of the 24 experimental design cells, four fit indices (AIC, BIC, AICc, CAIC) were calculated for each of the four model specifications, and model selection was conducted by comparing the values of each index across the four models. That study revealed that higher between-dimension correlations and smaller sample sizes influence the accuracy with which each index indicated the correct two-dimension generating structure. The four indices had 100% accuracy when between-dimension correlation was between .00 and .60. When between-dimension correlation was .75, accuracy was only less than optimal at N=100. Of the four indices, when between-dimension correlation was very high (.90), AIC performed better at

N=100 than the other three in terms of indicating the correct two-dimension generating structure, and at larger sample sizes AIC and AICc perform similarly and better than CAIC and BIC.

Following up on their initial study, Harrell and Wolfe (2009b) included two more independent variables in a more comprehensive study of global fit indices in multidimensional analysis of dichotomous data. Based on the findings that between-correlation of .75 and greater resulted in decreased accuracy in the indices, the second study focused on correlations of .6, .7, .8, and .9. As accuracy had not been an issue at a sample size of 1000, this study used sample sizes of 100, 250, 500 and 750. The two new variables were test length (20 and 40 items) and model-to-data misfit. Misfit was introduced by simulating data from multidimensional 1 parameter (M1PL), 2 parameter (M2PL) and 3 parameter (MC3PL) IRT models. The results showed that higher values of between-dimension correlation lead BIC and CAIC to indicate the correct two-dimension generating structure with slightly less frequency than AIC. However, with correlations greater than .70, none of the indices perform optimally. BIC and CAIC tends to indicate the one-dimension model, while AIC indicates the three-dimension model when the incorrect model is chosen. In addition, these results demonstrated that violations of the Rasch model assumptions are magnified at higher between-dimension correlations.

The current study builds on the work of Kang and Cohen (2007) and Harrell and Wolfe (2009a; 2009b) to further investigate the use of the global fit indices as applied to CFA in an item response theory framework. Our focus is on the accuracy of decisions in a multidimensional Rasch analysis of polytomous data using four global fit indices as a function of sample size, number of test items, data-to-model misfit, and between-dimension correlation. It is hypothesized that smaller sample sizes, shorter tests, higher levels of between-dimension

correlation will result in decreases in accuracy. It is also anticipated that higher levels of data-to-model misfit will decrease the accuracy of the indices.

Methods

Simulation Design

Data was simulated using ConQuest to represent polytomous item responses according to a two-dimensional rating scale model. The rating scale version of the MRCMLM holds the threshold parameters, or steps between the response categories, constant, in comparison to a partial credit model where the threshold parameters are allowed to vary for each item. The thresholds were selected as to conform to Linacre's essential criteria (2004). The thresholds increase in value as the categories progress. The distribution of the rating scale responses is unimodal, with at least 10 observations per response category.

Examinee abilities were generated from a bivariate standard normal distribution. Item difficulty generating parameters were drawn from a standard normal distribution. Between-dimension ***correlation*** (.60, .70, .80, & .90) served as one independent variable. These values were chosen based on Harrell and Wolfe (2009a), which found that the accuracy of the fit indices was not affected at values of .60 or lower. ***Sample size*** (100, 250, 500, & 750) served as the second independent variable. ***Test length*** (20 & 40) modeled shorter and moderate length tests. **Rating scale points** (3 point and 5 point rating scale) served as the fourth independent variable. The three point rating scale had thresholds of -0.67 and 0.67. The five point rating scale had thresholds of -1.8, -0.60, 0.60, and 1.8. There were 500 replications for each cell of the 64 cells of the [4 (correlation) x 4 (sample size) X 2 (test length) X 2 (rating scale structure)] experimental design. In all of the cases, two dimensions were simulated with half of the items defining each dimension.

Analysis

Each simulated data set was scaled to four configurations of the MRCMLM using ConQuest (Wu, Adams, Wilson, & Heldane, 2007). The configurations represented the following types of model specification: (a) under specification (i.e., a unidimensional model), (b) correct specification (i.e., a two-dimensional model with correct item-to-dimension mapping), (c) mis-specification (i.e., a two-dimensional model with incorrect item-to-dimension mapping), and (d) over specification (i.e., a three-dimensional model). Table 4.1 represents the mappings of the items for each of these model specifications for two test lengths.

Table 4.1. Item-to-dimension mappings

Test Length	Specification	Number of Dimensions	Items in Dimension 1	Items in Dimension 2	Items in Dimension 3
20 Items	Underfit	1	1-20		
	Correct	2	1-10	11-20	
	Misspecified	2	1-5, 16-20	6-15	
	Overfit	3	1-5	5-10	11-20
40 Items	Underfit	1	1-40		
	Correct	2	1-20	21-40	
	Misspecified	2	1-10, 31-40	11-30	
	Overfit	3	1-20	21-30	31-40

The values of AIC, AICc, BIC, and CAIC were calculated for each scaled data set using the log likelihood values produced by ConQuest. The minimum value of each fit index across the four model specifications shown in Table 4.1 was used as the selection criterion. The model satisfying the criterion for the index in question was designated as the "selected model" for each index for each simulated data set. The proportion of data sets for the 500 replications within each

cell of the experimental design that selected each of the 4 model specifications (i.e., the dependent variable) was computed and compared between experimental conditions.

A repeated measures logistic regression analysis was completed for each study to determine which effects were significant. The analysis modeled a dichotomous outcome, whether the index resulted in the correct or incorrect selection of the generating dimensionality structure, as a function of the main effects and two-way interactions of the independent variables and the indices. A repeated measures analysis was used since the same dataset was analyzed using each of the four indices. Odds ratios were used as indicators of effect sizes, with the cut-off for practical significance being an odds ratio greater than 2.0 (or less than 0.5, for negative relationships). This metric for a large effect size is common in medical sciences (Monahan, McHorney, Stump, & Perkins, 2007).

Results

Table 4.2 presents the results of a logistic regression analysis that contains significant two-way interaction terms crossing each of the independent variables with the global fit indices, as well as with the other independent variables. Three of the interaction effects had large effect sizes (i.e. odds ratios greater than 2 or less than .50): "correlation x index", "correlation x rating scale", and "correlation x number of items".

Table 4.2. Hypothesis Testing and Effect Sizes from Logistic Regression

Effect	Test Statistic	p-value	Odds Ratio
Main Effect			
Sample Size (N)	11.51	< .0001	1.01
Test Length (NI)	8.09	< .0001	3.70
Between-Dim. Correlation (CORR)	3.03	.0025	99061.88
Rating Scale Structure (RS)	9.15	.0340	5.81
Index: AIC	14.08	< .0001	977,740
Index: BIC	15.89	< .0001	13.57
Index: AICc	14.33	< .0001	43.68

Interaction Effects			
NI*AIC	13.83	< .0001	1.06
NI*AICC	-25.29	< .0001	.97
NI*BIC	11.84	< .0001	1.02
CORR*AIC	-15.11	< .0001	0
CORR*AICC	-12.84	< .0001	.26
CORR*BIC	-17.03	< .0001	.07
AIC*RS	-11.25	< .0001	.44
AICC*RS	-7.43	< .0001	.82
BIC*RS	-11.08	< .0001	.81
CORR*RS	-33.31	< .0001	.01
CORR*NI	-6.72	< .0001	.29

Note: The reference levels for the categorical variables in the model are not listed: the index CAIC and the interactions that include this term.

Figure 4.1 and Table 4.3 depict the average accuracy for each index at the four levels of between-dimension correlation. Although the accuracy of all the indices decline as the correlation increases, the differences among the performances of each index at the different levels changes. Figure 4.1 shows that there is more variation among the indices at correlations of .70 and .80 than at .60 and .90. AIC and AICc only achieve optimal performance (i.e. less than 5% error) at the between-dimension correlation of .60, while BIC and CAIC never achieve this level of accuracy. When the indices are not selecting the correct generating structure, they almost always indicated the one-dimension structure.

Figure 4.1. Correlation by Index Interaction Plot

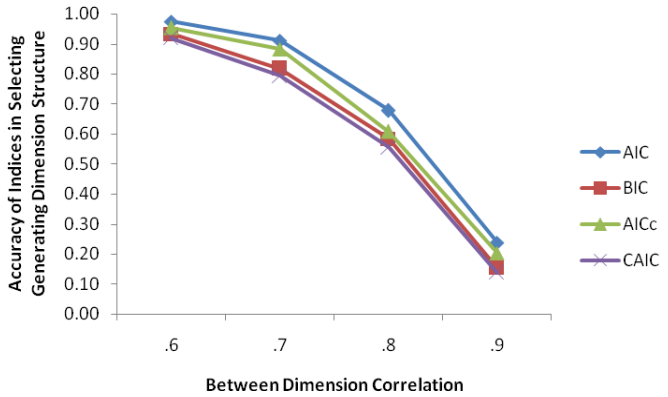

Table 4.3. Accuracy of Indices by Level of Correlation

	Between- Dimension Correlation															
	.60				.70				.80				.90			
Index	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3
AIC	.02	.97	**	0	.08	.91	**	**	.32	.67	**	0	.76	.23	**	**
BIC	.07	.93	0	0	.18	.82	0	0	.42	.58	0	0	.85	.15	0	0
AICc	.04	.95	**	0	.11	.88	**	0	.39	.60	**	0	.79	.20	**	**
CAIC	.08	.92	0	0	.21	.79	0	0	.44	.56	0	0	.86	.14	0	0

Note: 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure; ** indicates that the dimensional structure was indicated less than 1% of the time.

The correlation by rating scale interaction is associated with a large effect size. Figure 4.2 and Table 4.4 depict this interaction. While accuracy decreases for both the 3 point and 5 point rating scale data as the between-dimension correlation increases, accuracy decreases at a steeper rate for the five point rating scale. For the 5 point rating scale data with a between-dimension correlation of .90, the one-dimensional structure is indicated almost 100% of the time. Table 4.4 shows that the incorrect two-dimensional and three dimensional structure are indicated with frequencies less than 1% of the time.

Figure 4.2. Correlation by Rating Scale Interaction Plot

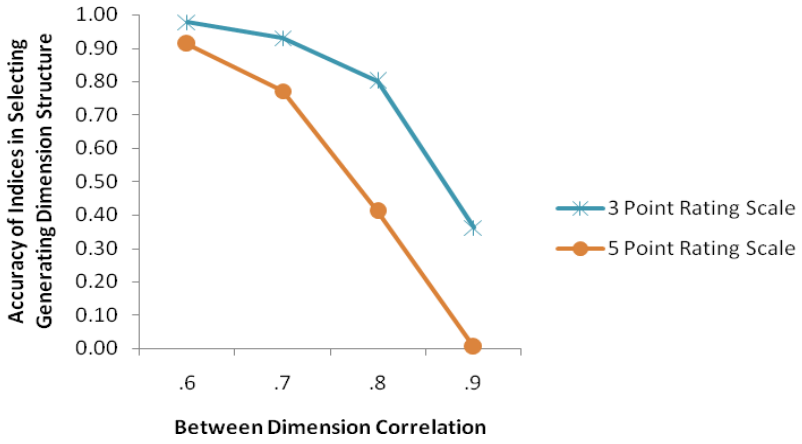

Table 4.4. Accuracy by Correlation and Rating Scale Structure

Rating Scale Structure	Between- Dimension Correlation															
	.60				.70				.80				.90			
	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3
3 Point Scale	.02	.98	0	0	.07	.92	**	**	.19	.80	**	0	.63	.36	**	**
5 Point Scale	.09	.91	0	0	.22	.77	**	0	.59	.41	0	0	.99	**	**	**

Note: 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure; ** indicates that the dimensional structure was indicated less than 1% of the time.

Figure 4.3 and Table 4.5 depict the interaction between test length and correlation. While accuracy decreases for both test lengths as the between-dimension correlation increases, accuracy for the 40 item test length is less affected by the change until the between-dimension correlation increases above .80. Accuracy for the 40 item test length does not negatively impact the accuracy of the indices at correlation levels of .60 or .70, while the average for the 20 item test length is never near the optimal level. For the 20 items tests with a between-dimension correlation of .90, the one-dimensional structure is indicated almost 100% of the time. Table 4.5 shows that the incorrect two-dimensional and three dimensional structure are indicated with frequencies less than 1% of the time.

Figure 4.3. Correlation by Test Length Interaction Plot

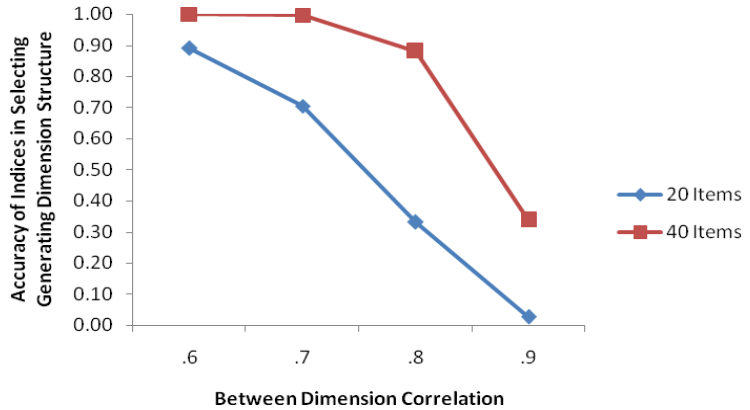

Table 4.5. Accuracy by Correlation and Test Length

Test Length	Between- Dimension Correlation .60				.70				.80				.90			
	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3	1	2C	2W	3
20 Items	.10	.89	**	0	.29	.70	**	**	.66	.33	**	0	.97	.02	**	**
40 Items	0	1	0	0	**	.99	0	0	.12	.88	0	0	.66	.33	.03	0

Note: 1= One dimensional model structure; 2C = Two dimensional correct (generating) model structure; 2W = Two dimensional mis-specified model structure; 3 = Three dimensional model structure; ** indicates that the dimensional structure was indicated less than 1% of the time.

Conclusions

Our results indicate that three two-way interactions are important when one considers the performance of global fit indices in dimensionality assessment in a multidimensional Rasch model application of confirmatory factor analysis. First, the index by correlation interaction indicates that higher values of between-dimension correlation affect the accuracy of BIC and CAIC more than the other indices, as AIC and AICc indicate the generating model more frequently at higher correlations. Unfortunately, even the performance of AIC is still not at or above optimal levels of performance so the use of AIC does not guarantee the accuracy of dimensionality decisions. When the indices do not select the generating dimension structure, they indicate the one-dimension model instead. In reality, it is commonplace for a one-dimensional model to be used when correlation between two dimensions is .90. However, the indices exhibit this same behavior at correlations of .7 and .8 more than 5% of the time. Generally, correlations of this size do not lead to the decision to treat data as unidimensional.

While the previous work by Harrell and Wolfe (2009a, 2009b) indicates that the accuracy of BIC is more strongly affected by higher levels of between-dimension correlation than AIC, the behavior of the indices when not indicating the correct generating dimensional structure is different. Both the initial study of dichotomous data (2009a) and this study indicate that the one-dimensional model was more likely to be selected than the three-dimensional structure or the

incorrect two-dimensional study. For dichotomous data, the three-dimensional structure was indicated by AIC and AICc more frequently than by BIC or CAIC. The initial dichotomous study (2009a) used a different item-to-dimension mapping than what was used in the second dichotomous study (2009b) and the current study.

Second, the results associated with the correlation and test length interaction is consistent with expectations. The combination of shorter tests and higher correlations results in a difficult-to-detect distinction being modeled with less statistical information. Previous research has shown that test length is a factor in the recovery of model parameters, especially with higher between-dimension correlations (Bolt and Lall, 2003; Kang & Cohen, 2007). However, Harrell and Wolfe (2009b) did not find as strong of a relationship with dichotomous data as was observed in the current study, which focused on polytomous data.

Third, our results indicated that as between-dimension correlation increased, accuracy was higher for data with the fewer rating scale categories. This finding is cause for concern as most instruments employ a rating scale with more than three categories. In fact, Wolfe and Smith (2007) recommended between three and five categories in a rating scale. Since dichotomous data is just polytomous data with only two categories, an analysis of the dichotomous data in Harrell and Wolfe (2009b) and the polytomous study in terms of rating scale data with two, three and five categories could shed some light on this result. Such an analysis would benefit from treating the difference in the number of scale categories as differences in total scores.

One limitation of the current study is our use of one set of thresholds for each of the rating scales. It is possible that the thresholds that are further apart or closer together could influence the generalizability of our results. In addition, we both generated data and recovered parameters using a Rasch rating scale model. Real data that more closely follows a partial credit

model may produce results that differ from ours as well as would data that follows a generalized rating scale or partial credit (i.e., two parameter logistic) model. The effect of these variables (thresholds spacing, rating scale/partial credit model, and model-data misspecification) on accuracy of dimensionality decisions that are based on global fit indices in applications of confirmatory factor analysis on the context of multidimensional Rasch analysis warrants further investigation.

References

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov&F. Csaki (Eds.), Second international symposium on information theory (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317–332.

Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28,* 351–374.

Bolt, D.M. & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 29, 395-414.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Harrell, L. M., & Wolfe, E.W. (2009a). *Effect of between-dimension correlation and sample size on multidimensional Rasch analysis.* Paper to be presented at annual meeting of the American Educational Research Association, San Diego, CA.

Harrell, L. M., & Wolfe, E.W. (2009b). *The Influence of Misfit on Multidimensional Rasch Model Global Fit Index Accuracy.* Manuscript in unpublished doctoral dissertation. Virginia Polytechnic Institute and State University, Department of Educational Leadership, Educational Research and Evaluation Program.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Communication in Statistics. Part B: Simulation & Computation, 26,* 1477–1516.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.

Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika,* 76 (2), 297-307.

Kang, T. & Cohen, A.S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement,* 31 (4), 331-358.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research,* 33 (2), 188-229.

Linacre, J. M. (2004). Optimizing Rating Scale Effectiveness. In E.V. Smith & R. M. Smith, Introduction to Rasch Measurement (pp.258-278). Maple Grove, MN: JAM Press.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

McQuarrie, A. D. R. & Tsai, C.L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

Monahan, P. O., McHorney, C. A., Stump, T.E., & Perkins, A.J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics,* 32 (1), 92-109.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6 (2), 461-464.

Weakliem, D.L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research,* 33 (2), 167-187.

Whittaker, T. A. & Stapleton, L. M. (2006). The performance of cross-validation indices used to select competing covariance structure models under multivariate nonnormality conditions. *Multivariate Behavioral Research,* 41 (3), 295-335.

Wolfe, E.W., & Smith, E.V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—validation activities. *Journal of Applied Measurement*, *8*, 204-234.

Wu, M. L., Adams, R. J., Wilson, M. R. , & Heldane, S.A. (2007). ACER ConQuest: Generalized item response modeling software (Version 2.0) [computer software]. Melbourne: Australian Council for Educational Research.

CHAPTER 5

DISCUSSION

Summary of Findings

Table 5.1 summarizes the results of the regression analyses conducted in the two studies. Interactions denoted as N/A were not included in the final logistic regression model because (a) they were not in an independent variable in that study or (b) they were not statistically significant. Only one of the common index-independent variable interactions was statistically significant and practically important in both studies: the index-correlation interaction. Of the common independent variable interactions, none were significant and practically important in both studies. The correlation by test length interaction had an odds ratio indicating practical importance in the polytomous study but not the dichotomous study.

Table 5.1. Interaction Effects with Practical Significance

Interaction	Dichtomous Data Study	Polytomous Data Study
Index - Correlation	**	**
Index - Sample Size	NS	N/A
Index - Test Length	NS	NS
Index - IRT Generating Model	NS	N/A
Index - Rating Scale Structure	N/A	NS
Correlation - Sample Size	NS	NS
Correlation - Test Length	NS	**
Correlation - IRT Generating Model	**	N/A
Correlation – Rating Scale Structure	N/A	**
Sample Size – Test Length	NS	N/A
Index – IRT Generating Model	NS	N/A

Note: NS=Not Practically Significant (odds ratio between .5 and 2); N/A=term not in model; ** Practical Significance (odds ratio less than .5 or greater than 2);

The interaction effect between global fit index and between-dimension correlation had very large effect sizes in both studies. At higher values of between-dimension correlation, AIC

indicates the correct two-dimension generating structure slightly more often than does the BIC or CAIC. However, the pattern of incorrect responses differed between the two studies. Both the initial study of dichotomous data (2009a) and the polytomous study indicate that the one-dimensional model was more likely to be selected than the three-dimensional structure or the incorrect two-dimensional study. For dichotomous data, the three-dimensional structure was indicated by AIC and AICc more frequently than by BIC or CAIC. The item-to-dimension mapping for the three-dimensional structure was different in the initial dichotomous study (2009a), as compared to the dichotomous and polytomous studies in this dissertation so the difference in structure does not explain the results. One possible reason may have something to do with data-to-model misfit. In the initial dichotomous study (2009a) and the polytomous study, the data was generated by the MRCMLM and scaled to the same model, while in the data in the dichotomous study in the dissertation was generated by a one-,two-,or three-parameter MIRT model and scaled to the MRCMLM.

The correlation by test length interaction had an odds ratio indicating practical importance in the polytomous study but not the dichotomous study. The direction of this relationship in the polytomous study is consistent with expectations. The combination of shorter tests and higher correlations results in a difficult-to-detect distinction being modeled with less statistical information. A possible explanation for the discrepancy in results could be due to the more complex nature of modeling polytomous data, as compared to modeling dichotomous.

The correlation by rating scale structure interaction also had an odds ratio indicating practical importance in the polytomous study. As between-dimension correlation increased, accuracy was higher for data with the fewer rating scale categories. Since dichotomous data is just polytomous data with only two categories, an analysis of the dichotomous data in Harrell &

Wolfe (2009b) and the polytomous study in terms of rating scale data with two, three and five categories could shed some light on this result.

In terms of the dichotomous study, only one other interaction other than the correlation by index interaction was had an odds ratio indicating practical importance. The correlation by IRT generating model term represented the interaction of between-dimension correlation and data-to-model misfit. As expected, the results demonstrate that violations of the Rasch model assumptions are magnified at higher between-dimension correlations.

In summary, there are a few recommendations for practitioners if they are working with highly correlated multidimensional data. When creating instruments, moderate lengths (roughly 40 items) should be used instead of shorter lengths (such as 20 items). The model used for confirmatory factor analysis (MRCMLM or other MIRT models) should be chosen to minimize data-to-model misfit. Finally, decisions about dimensionality should be based on multiple global indices instead of depending on one index in particular.

Future Directions for Research

In terms of dichotomous data, further research is warranted in terms of the way that Rasch-violations are modeled in simulation studies such as this. In our study, the two and three-parameter models were generated with a reasonable amount of variability in the slope violation. Further research simulating a narrower range of slopes should be investigated to further explore this finding. In addition, our results can be extended by examining the accuracy of dimensionality decisions based upon global fit indices when data are scaled to the M2PL and MC3PL.

One limitation of the polytomous study is our use of one set of thresholds for each of the rating scales. It is possible that the thresholds that are farther apart or closer together could influence the generalizability of our results. In addition, we both generated data and recovered parameters using a Rasch rating scale model. Real data that more closely follows a partial credit model may produce results that differ from ours as well as would data that follows a generalized rating scale or partial credit (i.e., two parameter logistic) model. The effect of these variables (thresholds spacing, rating scale/partial credit model, and model-data misspecification) on accuracy of dimensionality decisions that are based on global fit indices in applications of confirmatory factor analysis on the context of multidimensional Rasch analysis of polytomous data warrants further investigation.

Additionally, the work from these studies could be link the research on global fit indices in the structural equation modeling (SEM) framework. One such study could compare the agreement of dimensionality decisions across the frameworks using AIC and BIC. Another would be to incorporate other indices used SEM, such as root mean square error of approximation (RMSEA) and Comparative Fit Index (CFI), into the multidimensional Rasch analysis framework.

Some additional work can be done with the existing data. The trend that was identified in Harrell and Wolfe (2009a), the sample size by correlation interaction, was not found to be practically significant in either of the studies here. An analysis of the variables in the two studies to determine if multicollinearity or another phenomenon played a part in this result would be helpful. Additionally, the previously mentioned combined analysis of the dichotomous and polytomous data in terms of two, three and five category rating scale structures should be

conducted. Such an analysis would benefit from treating the difference in the number of scale categories as differences in total scores.

Learning from the Dissertation Process

The process of completing this dissertation was a learning experience in many ways. Prior to the analysis conducted in these studies, I had never used repeated measures logistic regression. In fact, my use of Proc Genmod in SAS had been limited to homework assignments during coursework for my M.S. in Statistics at the University of Georgia. Multisim, which I used to generate data in the dichotomous study, was new software that I had to learn as well. In addition, I had never seen or used the two- or three-parameter multidimensional IRT models.

The process of turning the apprenticeship project into two dissertation studies and identifying future research questions from the results of dissertation studies has been a valuable lesson as well. This is unlike picking topics for projects for courses that are often unrelated or hard to extend to another course. Through the completion of the apprenticeship and dissertation processes, I have come across other questions based on the results I have obtained that will serve as the basis for my initial research agenda regarding the use of global fit indices in multidimensional Rasch analysis. This is valuable to me as learning to develop a continuing base for research is necessary for succeeding as a researcher in academia or private industry.

When I was earning my masters, I did not believe that I would want to be in a research-based profession. The various research experiences during my doctoral study, including the dissertation, have changed that point of view. In addition, completing the dissertation process has also made me confident that I am capable of doing quality research.

REFERENCES

(Not included in Chapter 3 or 4)

Anderson, D. R., Burnham, K.P., & White, G.C. (1998). Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics,* 25(2), 263-282.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement,* 4, 87-100.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research,* 33(2), 261-304.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. VanderLinden & R. K. Hambleton (Eds.), Handbook *of modern item response theory* (pp. 271-296). New York: Springer-Verlag.

Spray, J. A., Davey, D. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). Comparison of two logistic multidimensional item response theory models (ACT research report series ONR90-8). Iowa-City, IA: ACT Inc.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.

Wang, C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and Psychological Measurement*, 64(6), 937-955.

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

Yang, C.C., & Yang, C.C. (2007). Separating latent classes by information criteria. *Journal of Classification,* 24, 183-203.

APPENDIX A

SAMPLE DATA GENERATION SAS CODE FOR DICHOTOMOUS DATA STUDY

This code uses SAS to create data in Multisim, scale the data to four different dimension structures in Conquest, and create cumulative output files for analysis purposes.

```
options nodate nonumber macrogen mprint nospool xsync noxwait nocenter;
%macro datagen(corr,n,ni,iter,start_seed,gen_model);
%do iter=1 %to &iter.;

%let seed=%eval((&iter.*&gen_model.)+&n.+&start_seed.);

***create file for Multisim code and runs Multisim***;
*************************************************;

*create file for Conquest code;
data one; ni=&ni.; file 'c:\Multisim\msim.in'; put "'o&iter.&n..txt'" / '2' /
ni;
 run;

*generate two correlated dimensions;
data items;
var=.25;
 do i=1 to 10;
  X1 = sqrt(var)*RanNor(&Seed.);
  a1 = exp(x1);
  a2=0;
  a3=0;
  a4=0;
   b=rannor(&seed.);
  d = -1*a1*b;
  c=.2;
  output;
 end;

 do j=1 to 10;
  a1=0;
  X2 = sqrt(var)*RanNor(&Seed.);
 a2 = exp(x2);
  a3=0;
  a4=0;
 b=rannor(&seed.);
  d = -1*a2*b;
  c=.2;
  output;
  end;
run;

*writes data to file;
data two;
 file 'c:\Multisim\msim.in' mod;
 set items;
```

```
 put a1 a2 a3 a4 d c;
run;

*writes Multisim code to file;
data three;
corr = &corr.;
iter=&iter.;
sampsize=&n.;
seed = -1*&seed.;
 file 'c:\Multisim\msim.in' mod;
 put sampsize / '1' / '1' / '0' / '0' /
 '2' / '0' / '0' / '0' / '0' / corr / '0' / '0' / '0' / '0' / '0' /
  seed / '1' / "'d&iter.&n..txt'";
run;

* calls Multisim through SAS;
x "cd c:\multisim";
x "c:\multisim\msim";


***create and run Conquest code***;
*********************************;

*creates the four dimensionality structure files;
data struc_gen_1;
file 'C:\multisim\md_struc_1.cqc';
 put "datafile d&iter.&n..txt;" / "format id 1-4 responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1) (0,1)  !items(1-20);"
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> onedim1.out;"/
"show !tables=2 >> onedim2.out;"/
"show !tables=3 >> onedim3.out;"/
"show  cases ! estimates=latent >> onedimest.out;"/
"quit;";
run;

data struc_gen_2c;
file 'C:\multisim\md_struc_2c.cqc';
 put "datafile d&iter.&n..txt;" / "format id 1-4 responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1) (0,1) (   ) !items(1-10);" /
 " score (0,1) (   ) (0,1) !items(11-20);"/
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> twodimc1.out;"/
"show !tables=2 >> twodimc2.out;"/
"show !tables=3 >> twodimc3.out;"/
"show  cases ! estimates=latent >> twodimcest.out;"/
"quit;";
run;

data struc_gen_2w;
file 'C:\multisim\md_struc_2w.cqc';
 put "datafile d&iter.&n..txt;" / "format id 1-4 responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1) (0,1) (   ) !items(1-5,16-20);" /
 " score (0,1) (   ) (0,1) !items(6-15);"/
```

```
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> twodimw1.out;"/
"show !tables=2 >> twodimw2.out;"/
"show !tables=3 >> twodimw3.out;"/
"show  cases ! estimates=latent >> twodimwest.out;"/
"quit;";
run;

data struc_gen_3;
file 'C:\multisim\md_struc_3.cqc';
 put "datafile d&iter.&n..txt;" / "format id 1-4 responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1) (0,1) (   ) (   ) !items(1-5);"/
"score (0,1) (   ) (0,1) (   ) !items(6-15);"/
"score (0,1) (   ) (   ) (0,1) !items(16-20);"/
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> threedim1.out;"/
"show !tables=2 >> threedim2.out;"/
"show !tables=3 >> threedim3.out;"/
"show  cases ! estimates=latent >> threedimest.out;"/
"quit;";
run;

*runs the above Conquest code;
x console md_struc_1.cqc;
x console md_struc_2c.cqc;
x console md_struc_2w.cqc;
x console md_struc_3.cqc;


*read dim analyses output from Conquest files to
create a cumulative file for each of the four dimensional structures;
*onedim;
data cumulative1;
 infile 'C:\multisim\onedim1.out' truncover scanover;
 input @'The Data File:' in_file $char14. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_1 12.6
@'Total number of estimated parameters:' no_param_1 4.0 @'The number of
iterations:' n_iter 4.0
@ 'Iterations terminated' term_com $char70.;
run;

data cumulative1;
 file 'C:\multisim\onedimcum.txt' mod;
 set cumulative1;
 it=&iter.;
 corr=&corr.;
 n=&n.;
 put n samp_sze dev_1 no_param_1 in_file it corr n_iter term_com;
run;

DATA person1d;
  Infile 'C:\multisim\onedimest.out';
  Input itemnum / / / / / / estdim1 / sedim1;
RUN;

data person1d;
```

```
 file 'C:\multisim\onedimpers.txt' mod;
 set person1d;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 sedim1;
run;

data itemest1d;
infile 'C:\multisim\onedim2.out' truncover missover;
input /////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest1d;
 file 'C:\multisim\onedimitemest.txt' mod;
 set itemest1d;
iter=&iter.;
samp_sze=&n.;
corr=&corr.;
* iter=1;
 *samp_sze=2;
 *corr=3;
 put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

 *two dimension correct;
data cumulative2c;
 infile 'C:\multisim\twodimc1.out' truncover scanover;
 input @'The Data File:' in_file $char14. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_2c 12.6  @'Total number of estimated parameters:'
no_param_2c 4.0
@'The number of iterations:' n_iter 4.0 @ 'Iterations terminated' term_com
$char70.;
run;

data cumulative2c;
 file 'C:\multisim\twodimcumc.txt' mod;
 set cumulative2c;
it=&iter.;
corr=&corr.;
n=&n.;
put n samp_sze dev_2c no_param_2c in_file it corr n_iter term_com;
run;
```

```
DATA person2c;
  Infile 'C:\multisim\twodimcest.out';
  Input itemnum / / / / / / estdim1 estdim2 / sedim1 sedim2;
RUN;

data person2c;
 file 'C:\multisim\twodimcpers.txt' mod;
 set person2c;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 sedim1 sedim2;
run;

DATA corr2c;
      infile 'C:\multisim\twodimc3.out';
      input ////////////////////// dimname $26. correst 6.2 @ ;
run;

data corr2c;
      file 'C:\multisim\twodimccorr.txt' mod;
set corr2c;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr correst;
run;


data itemest2c;
infile 'C:\multisim\twodimc2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest2c;
 file 'C:\multisim\twodimcitemest.txt' mod;
 set itemest2c;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

 *two dimension wrong;
data cumulative2w;
```

```
 infile 'C:\multisim\twodimw1.out' truncover scanover;
 input @'The Data File:' in_file $char14. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_2w 12.6  @'Total number of estimated parameters:'
no_param_2w 4.0
@'The number of iterations:' n_iter 4.0 @ 'Iterations terminated' term_com
$char70.;
run;

data cumulative2w;
 file 'C:\multisim\twodimwumc.txt' mod;
 set cumulative2w;
it=&iter.;
corr=&corr.;
n=&n.;
 put n samp_sze dev_2w no_param_2w in_file it corr n_iter term_com;
run;

DATA person2w;
  Infile 'C:\multisim\twodimwest.out';
  Input itemnum / / / / / / estdim1 estdim2 / sedim1 sedim2;
RUN;

data person2w;
 file 'C:\multisim\twodimwpers.txt' mod;
 set person2w;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 sedim1 sedim2;
run;

DATA corr2w;
      infile 'C:\multisim\twodimw3.out';
      input //////////////////// dimname $26. correst 6.2 @ ;
run;

data corr2w;
      file 'C:\multisim\twodimwcorr.txt' mod;
set corr2w;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr correst;
run;


data itemest2w;
infile 'C:\multisim\twodimw2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
```

```
;
run;

data itemest2w;
 file 'C:\multisim\twodimwitemest.txt' mod;
 set itemest2w;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

*three dimension;
data cumulative3;
 infile 'C:\multisim\threedim1.out' truncover scanover;
 input @'The Data File:' in_file $char14. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_3 12.6
@'Total number of estimated parameters:' no_param_3 4.0 @'The number of
iterations:' n_iter 4.0
@ 'Iterations terminated' term_com $char70.;
run;

data cumulative3;
 file 'C:\multisim\threedimcum.txt' mod;
 set cumulative3;
 it=&iter.;
 corr=&corr.;
 n=&n.;
 put n samp_sze dev_3 no_param_3 in_file it corr n_iter term_com;
run;

DATA person3d;
  Infile 'C:\multisim\threedimest.out';
  Input itemnum / / / / / / estdim1 estdim2 estdim3 / sedim1 sedim2 sedim3;
RUN;

data person3d;
 file 'C:\multisim\threedimpers.txt' mod;
 set person3d;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 estdim3 sedim1 sedim2 sedim3;
run;

DATA corr3d;
      infile 'C:\multisim\threedim3.out';
      input ////////////////////// dimname $26. correst1_2 6.2 / dimname2 $26.
correst1_3 6.2 correst2_3 8.2 @@ ;
run;

data corr3d;
      file 'C:\multisim\threedimcorr.txt' mod;
set corr3d;
```

```
 iter=&iter.;
 samp_sze=&n.;
corr=&corr.;
 put samp_sze iter corr correst1_2 correst1_3 correst2_3;
run;

data itemest3d;
infile 'C:\multisim\threedim2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest3d;
 file 'C:\multisim\threedimitemest.txt' mod;
 set itemest3d;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

%end;
%mend datagen;
*%macro datagen(corr,n,ni,iter,start_seed,gen_model);

*runs macro;
%datagen(.9,750,20,200,99200,3);
```

APPENDIX B

SAMPLE DATA GENERATION SAS CODE FOR POLYTOMOUS DATA STUDY

This code uses SAS to create data in Conquest, scale the data to four different dimension structures in Conquest, and create cumulative output files for analysis purposes.

```
*max seed 2,147,483,647;
options nodate nonumber macrogen mprint nospool xsync noxwait nocenter;
%macro datagen(corr,n,ni,ndim,niter,start_seed,nthresh);

%do iter=1 %to &niter.;

%let seed=%eval((&iter.*&nthresh.+&n.+&start_seed.);

*generate Conquest dimensionality structure files;
data struc_gen_1;
file 'C:\multisim\md_struc_1.cqc';
 put "datafile d_&nthresh._&ni._&n._&corr._&iter..txt;" / "format id 1-4
responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1,2,3,4) (0,1,2,3,4) !items (1-20); "
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> onedim1.out;"/
"show !tables=2 >> onedim2.out;"/
"show !tables=3 >> onedim3.out;"/
"show  cases ! estimates=latent >> onedimest.out;"/
"quit;";
run;

data struc_gen_2c;
file 'C:\multisim\md_struc_2c.cqc';
 put "datafile d_&nthresh._&ni._&n._&corr._&iter..txt;" / "format id 1-4
responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1,2,3,4) (0,1,2,3,4) () !items(1-10);" /
 " score (0,1,2,3,4) () (0,1,2,3,4) !items(11-20);"/
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> twodimc1.out;"/
"show !tables=2 >> twodimc2.out;"/
"show !tables=3 >> twodimc3.out;"/
"show  cases ! estimates=latent >> twodimcest.out;"/
"quit;";
run;

data struc_gen_2w;
file 'C:\multisim\md_struc_2w.cqc';
 put "datafile d_&nthresh._&ni._&n._&corr._&iter..txt;" / "format id 1-4
responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1,2,3,4) (0,1,2,3,4) () !items(1-5,16-20);" /
 " score (0,1,2,3,4) () (0,1,2,3,4) !items(6-15);"/
"model item;"// "estimate ! method=gauss;"/
```

```
"show !tables=1 >> twodimw1.out;"/
"show !tables=2 >> twodimw2.out;"/
"show !tables=3 >> twodimw3.out;"/
"show !tables=4 >> twodimw4.out;"/
"show  cases ! estimates=latent >> twodimwest.out;"/
"quit;";
run;

data struc_gen_3;
file 'C:\multisim\md_struc_3.cqc';
 put "datafile d_&nthresh._&ni._&n._&corr._&iter..txt;"/ "format id 1-4
responses 1-20;"//
 "set constraints=cases,update=yes,warnings=no;"/
 "score (0,1,2,3,4) (0,1,2,3,4) (   ) (   ) !items(1-5);"/
"score (0,1,2,3,4) (   ) (0,1,2,3,4) (   ) !items(6-15);"/
"score (0,1,2,3,4) (   ) (   ) (0,1,2,3,4) !items(16-20);"/
"model item;"// "estimate ! method=gauss;"/
"show !tables=1 >> threedim1.out;"/
"show !tables=2 >> threedim2.out;"/
"show !tables=3 >> threedim3.out;"/
"show  cases ! estimates=latent >> threedimest.out;"/
"quit;";
run;

*generate difficulty parameters for items and put those and threshold into a
file;
data items;
 file 'c:\Multisim\trial.dat';
 do i=1 to &ni.;
   b=rannor(&seed.);
   format b 4.2;
   output;
   put  i "0 " b / i "1 -1.8" / i "2 -.6" /i "3 .6" / ;
 end;
run;

*generate Conquest command file;
data cmdflegen;
 file 'c:\Multisim\md_data.cqc';
 put "set seed= &seed.;"/ "set warnings = no;"/
/ "generate !nitems=&ndim.:&ndim., npersons=&n. ,
maxscore=4, itemdist=trial.dat," /
  "abilitydist=MVNORMAL(0:1:0:1:&corr.) >>
d_&nthresh._&ni._&n._&corr._&iter..txt;" /
  "quit;";
run;

*run 4 dimensionality analyses in Conquest on each dataset generated;
x console md_data.cqc;
x console md_struc_1.cqc;
x console md_struc_2c.cqc;
x console md_struc_2w.cqc;
x console md_struc_3.cqc;

*read dim analyses to create a cumulative file for each dim;
*onedim;
data cumulative1;
```

```
 infile 'C:\multisim\onedim1.out' truncover scanover;
 input @'The Data File:' in_file $char20. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_1 12.6
@'Total number of estimated parameters:' no_param_1 4.0 @'The number of
iterations:' n_iter 4.0
@ 'Iterations terminated' term_com $char70.;
run;

data cumulative1;
 file 'C:\multisim\onedimcum.txt' mod;
 set cumulative1;
 it=&iter.;
 corr=&corr.;
 n=&n.;
 ni=&ni.;
 nthresh=&nthresh.;
put n samp_sze ni corr nthresh it no_param_1 dev_1 n_iter in_file term_com;
run;

DATA person1d;
  Infile 'C:\multisim\onedimest.out';
  Input itemnum / / / / / / estdim1 / sedim1;
RUN;

data person1d;
 file 'C:\multisim\onedimpers.txt' mod;
 set person1d;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 sedim1;
run;

data itemest1d;
infile 'C:\multisim\onedim2.out' truncover missover;
input /////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest1d;
 file 'C:\multisim\onedimitemest.txt' mod;
 set itemest1d;
iter=&iter.;
samp_sze=&n.;
corr=&corr.;
* iter=1;
 *samp_sze=2;
 *corr=3;
```

```
 put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

 *twodim correct dimensionality structure;
data cumulative2c;
 infile 'C:\multisim\twodimc1.out' truncover scanover;
 input @'The Data File:' in_file $char20. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_2c 12.6  @'Total number of estimated parameters:'
no_param_2c 4.0
@'The number of iterations:' n_iter 4.0 @ 'Iterations terminated' term_com
$char70.;
run;

data cumulative2c;
 file 'C:\multisim\twodimcumc.txt' mod;
 set cumulative2c;
it=&iter.;
corr=&corr.;
n=&n.;
ni=&ni.;
 nthresh=&nthresh.;
put n samp_sze ni corr nthresh it no_param_2c dev_2c n_iter in_file term_com;
run;

DATA person2c;
  Infile 'C:\multisim\twodimcest.out';
  Input itemnum / / / / / / estdim1 estdim2 / sedim1 sedim2;
RUN;

data person2c;
 file 'C:\multisim\twodimcpers.txt' mod;
 set person2c;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 sedim1 sedim2;
run;

DATA corr2c;
      infile 'C:\multisim\twodimc3.out';
      input ////////////////////// dimname $26. correst 6.2 @ ;
run;

data corr2c;
      file 'C:\multisim\twodimccorr.txt' mod;
set corr2c;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr correst;
run;


data itemest2c;
```

```
infile 'C:\multisim\twodimc2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest2c;
 file 'C:\multisim\twodimcitemest.txt' mod;
 set itemest2c;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20 iter corr samp_sze;
 run;

 *two wrong dimensionality structure;
data cumulative2w;
 infile 'C:\multisim\twodimw1.out' truncover scanover;
 input @'The Data File:' in_file $char20. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_2w 12.6  @'Total number of estimated parameters:'
no_param_2w 4.0
@'The number of iterations:' n_iter 4.0 @ 'Iterations terminated' term_com
$char70.;
run;

data cumulative2w;
 file 'C:\multisim\twodimwumc.txt' mod;
 set cumulative2w;
it=&iter.;
corr=&corr.;
n=&n.;
ni=&ni.;
 nthresh=&nthresh.;
put n samp_sze ni corr nthresh it no_param_2w dev_2w n_iter in_file term_com;
run;

DATA person2w;
  Infile 'C:\multisim\twodimwest.out';
  Input itemnum / / / / / / estdim1 estdim2 / sedim1 sedim2;
RUN;

data person2w;
 file 'C:\multisim\twodimwpers.txt' mod;
 set person2w;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 sedim1 sedim2;
```

```
run;

DATA corr2w;
      infile 'C:\multisim\twodimw3.out';
      input ////////////////////// dimname $26. correst 6.2 @ ;
run;

data corr2w;
      file 'C:\multisim\twodimwcorr.txt' mod;
set corr2w;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr correst;
run;

data itemest2w;
infile 'C:\multisim\twodimw2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest2w;
 file 'C:\multisim\twodimwitemest.txt' mod;
 set itemest2w;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

*threedim dimensionality structure;
data cumulative3;
 infile 'C:\multisim\threedim1.out' truncover scanover;
 input @'The Data File:' in_file $char20. @'Sample size:' samp_sze 4.0
@'Final Deviance:' dev_3 12.6
@'Total number of estimated parameters:' no_param_3 4.0 @'The number of
iterations:' n_iter 4.0
@ 'Iterations terminated' term_com $char70.;
run;

data cumulative3;
 file 'C:\multisim\threedimcum.txt' mod;
 set cumulative3;
 it=&iter.;
 corr=&corr.;
```

```
 n=&n.;
 ni=&ni.;
  nthresh=&nthresh.;
put n samp_sze ni corr nthresh it no_param_3 dev_3 n_iter in_file term_com;
run;

DATA person3d;
  Infile 'C:\multisim\threedimest.out';
  Input itemnum / / / / / / estdim1 estdim2 estdim3 / sedim1 sedim2 sedim3;
RUN;

data person3d;
 file 'C:\multisim\threedimpers.txt' mod;
 set person3d;
 iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
 put samp_sze iter corr itemnum estdim1 estdim2 estdim3 sedim1 sedim2 sedim3;
run;

DATA corr3d;
      infile 'C:\multisim\threedim3.out';
      input ////////////////////// dimname $26. correst1_2 6.2 / dimname2 $26.
correst1_3 6.2 correst2_3 8.2 @@ ;
run;

data corr3d;
      file 'C:\multisim\threedimcorr.txt' mod;
set corr3d;
 iter=&iter.;
 samp_sze=&n.;
corr=&corr.;
 put samp_sze iter corr correst1_2 correst1_3 correst2_3;
run;

data itemest3d;
infile 'C:\multisim\threedim2.out' truncover missover;
input ////////// @6 in1 2.0 @22 i1 6.3 / @6 in2 2.0 @22 i2 6.3 / @6 in3 2.0
@22 i3 6.3 /@6 in4 2.0 @22 i4 6.3 /
@6 in5 2.0 @22 i5 6.3 /@6 in6 2.0 @22 i6 6.3 /  @6 in7 2.0 @22 i7 6.3 / @6
in8 2.0 @22 i8 6.3 / @6 in9 2.0 @22 i9 6.3
/ @6 in10 2.0 @22 i10 6.3 / @6 in11 2.0 @22 i11 6.3 / @6 in12 2.0 @22 i12 6.3
/ @6 in13 2.0 @22 i13 6.3
/@6 in14 2.0 @22 i14 6.3 /@6 in15 2.0 @22 i15 6.3 /@6 in16 2.0 @22 i16 6.3
/@6 in17 2.0 @22 i17 6.3 /@6 in18 2.0 @22 i18 6.3
/@6 in19 2.0 @22 i19 6.3 /@6 in20 2.0 @22 i20 6.3
;
run;

data itemest3d;
 file 'C:\multisim\threedimitemest.txt' mod;
 set itemest3d;
iter=&iter.;
 samp_sze=&n.;
 corr=&corr.;
put in1 i1 in2  i2 in3  i3 in4 i4 in5 i5 in6 i6 in7 i7 in8 i8  in9 i9 in10
i10 in11 i11 in12 i12 in13 i13 in14 i14
```

```
in15 i15 in16 i16 in17 i17 in18 i18 in19 i19 in20 i20
 iter corr samp_sze;
 run;

%end;
%mend datagen;
*%macro datagen(corr,n,ni,ndim,niter,start_seed,nthresh);

*runs macro;
%datagen(.70,500,20,10,200,80000,5);
```